



A machine learning approach to modelling social media sentiment for stock market trading in South Africa

J Steenkamp



[orcid.org/ 0000-0001-8652-2368](https://orcid.org/0000-0001-8652-2368)

Thesis accepted in fulfilment of the requirements for the degree *Doctor of Philosophy in Computer and Information Sciences with Computer Science and Information Systems* at the North-West University

Promoter: Dr A van der Merwe

Co-promoter: Prof HA Krüger

Graduation: May 2026

DEDICATION

This is dedicated to my father, my mother, my brother, my grandparents, and the family whose footsteps were laid long before mine. I carry their history in every part of my life, for I am built from their hands, courage, values, sacrifices, and lessons. Their lives echo through mine; it shapes the way I move through the world and the way I dream for the future. And, I suppose, to my cat (Blue), who spent as many hours sitting on my laptop as I was working on it. This work stands as a quiet tribute to the generations who built the foundation beneath me and to their enduring legacy that flows through me and continues to light the path in front of me.

ACKNOWLEDGEMENTS

I would not be where I am today without the support of some remarkable individuals. While I am grateful to everyone who has supported me along this journey, a few deserve special mention.

Firstly, to my study leaders, Doctor Annette and Professor Hennie - thank you. You once said we would ride the highs and lows together, but I was mainly on the highs. Your support, patience, and genuine interest have meant the world to me. You handled my confidence and my confusion equally well. Your influence will remain with me always. I am also grateful to my language editor, Mrs Elma van Wyk, who bravely and competently rehabilitated my Namibian English - also known as Namlish - into grammatically sound and academically acceptable English. I salute you.

To my colleagues and team members at FNB Namibia and Broader Africa, thank you for your flexibility and encouragement. It helped me balance work and academic responsibilities. I am incredibly thankful for the invaluable support and constant motivation throughout this journey.

For the friends closest to my heart, I want to acknowledge that your unwavering presence, reliability, and genuine sympathy have been truly invaluable. Thank you for pretending that I had it together at all. I am equally grateful to the families of my friends. Your kindness and support extended beyond friendship itself, offering guidance and reassurance when I needed it most.

To the extraordinary individuals who have stood by me throughout my life or during significant chapters of it - thank you for embracing me as your own. You gave me more than friendship; you gave me family. I am privileged to have you in my corner. I appreciate your constant support.

My deepest gratitude is extended to my family. You are my foundation. The daily love, support, and unwavering belief of my parents have been my guiding light. Their quiet strength continues to inspire me. I am grateful for a brother who is not only family but a lifelong friend. I also honour my grandparents, especially my grandmother who is still with us, whose presence and legacy continue to motivate and uplift me. Their remarkable legacy continues to inspire me in all I do.

I also want to take a moment to acknowledge the resilience, determination and, at times, patience I have demonstrated over the past few years. I am thankful for proving to myself that I can achieve anything I set my mind to. The output of this degree is not this thesis; it is me.

Lastly, but most important, I want to thank my God - for my gifts and the grace, honour and opportunity to use them. I am grateful for every answered prayer, the miracles disguised in the form of caffeine (occasionally wine), the quiet nudges hidden in the noise of chaos, and the gentle reminders that He was in control even when I clearly was not. I owe everything to Him.

ABSTRACT

The influence of sentiment and emotions on financial markets has increased markedly with the proliferation of social media data. However, empirical research on sentiment-driven trading strategies and their performance, particularly within the South African stock market context, remains limited. This study addresses this gap by developing a sentiment- and emotion-based trading framework, grounded in South African data and tailored to the Johannesburg Stock Exchange (JSE). The framework integrates Natural Language Processing (NLP) and machine learning techniques to model the relationship between media-derived sentiment and emotions and subsequent stock price movements.

A dataset of South African newspaper articles published between January 2018 and June 2024 was compiled and processed using lexicon-based sentiment scoring, in conjunction with supervised machine learning classifiers. Multiple pre-processing techniques and class balancing approaches were evaluated to address sentiment class imbalance. Stock price data for selected JSE-listed companies were collected, and correlation, lag, and position analyses were conducted to examine sentiment-price relationships. Hyperparameter tuning and optimisation techniques were applied to refine sentiment and emotion thresholds towards generating optimal trading signals. The Long Short-Term Memory (LSTM) model demonstrated superior performance, achieving robust classification results. The optimised sentiment and emotion thresholds were tested for the seven selected financial stocks and generated a hypothetical 26.07% average growth over the chosen testing period.

This research contributes to both academic and practical domains by presenting a trading framework specifically designed for an emerging South African market context. The findings demonstrate the potential of localised sentiment analysis in enhancing stock trading strategies and provide a methodological basis for integrating emotion-driven signals into market prediction models. The framework offers scope for further development using daily sentiment and emotion data to enhance predictive power and adaptability in dynamic market conditions.

Emotion analysis; Johannesburg Stock Exchange; Long Short-Term Memory model; machine learning; sentiment analysis; South Africa; trading strategy

TABLE OF CONTENTS

TITLE PAGE	i
DEDICATION	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
LIST OF FIGURES	x
LIST OF TABLES	xii
LIST OF ALGORITHMS	xiii
LIST OF ACRONYMS	xiii
Chapter 1 Introduction and problem contextualisation	15
1.1 Introduction	15
1.2 Problem statement	16
1.3 Secondary research objectives	18
1.4 Contributions	20
1.5 Scope and limitations	21
1.6 Structure of the thesis	24
1.6.1 Phase I – Introduction.....	24
1.6.2 Phase II – Business understanding.....	24
1.6.3 Phase III – Data understanding and preparation.....	25
1.6.4 Phase IV – Model development and evaluation	25
1.6.5 Phase V – Formulation and formalisation of the trading strategy	25
1.6.6 Phase VI – Conclusion	25
1.7 Conclusion	26
Chapter 2 The stock market and sentiment analysis	27
2.1 Introduction	27
2.2 The stock market	28
2.2.1 Overview of stock exchanges	29
2.2.2 The Johannesburg Stock Exchange	30
2.2.2.1 Historical development.....	30

2.2.2.2	Economic significance.....	31
2.2.2.3	Market structure and key indices.....	31
2.2.2.4	Market sectors	32
2.3	Stock price determination.....	33
2.4	Factors influencing stock prices.....	34
2.4.1	Macroeconomic factors.....	34
2.4.2	Microeconomic factors and firm-specific factors.....	35
2.4.3	Global events and market shocks	35
2.5	Sentiment analysis	36
2.5.1	Introduction.....	37
2.5.2	Sentiment analysis in the stock market.....	40
2.5.3	Sentiment analysis in the South African context	44
2.6	Machine learning	48
2.6.1	Introduction.....	48
2.6.2	Recent research in machine learning.....	49
2.7	Conclusion.....	52
Chapter 3	Sentiment analysis and machine learning.....	53
3.1	Introduction	53
3.2	Fundamentals of sentiment analysis	54
3.2.1	Definition and background	54
3.2.2	Data pre-processing	55
3.2.3	Analysis methodologies	58
3.3	Fundamentals of machine learning.....	61
3.3.1	Definition	61
3.3.2	Paradigms	61
3.3.3	Approaches	64
3.3.3.1	Shallow learning.....	65
3.3.3.1.1	Decision tree	65
3.3.3.1.2	The k-nearest neighbour algorithm.....	67
3.3.3.1.3	Naïve Bayes classifier	70
3.3.3.1.4	Support vector machine.....	71
3.3.3.1.5	Logistic regression	73
3.3.3.1.6	Random forests.....	76
3.3.3.2	Deep learning	78
3.3.3.2.1	Convolutional neural network	82

3.3.3.2.2	Recurrent neural network	86
3.3.3.2.2.1	Long Short-Term Memory model.....	87
3.4	Conclusion.....	89
Chapter 4	Methodology	90
4.1	Introduction	90
4.2	Data collection.....	91
4.2.1	Stock selection and historical stock prices	91
4.2.2	Stock media data	92
4.3	Data pre-processing.....	95
4.3.1	Feature extraction.....	95
4.3.2	Label assignment	96
4.4	Model selection	97
4.5	Model validation	100
4.5.1	Sentiment-driven machine learning model	100
4.5.2	Analytical methods for relationship analysis.....	104
4.5.2.1	Correlation analysis	105
4.5.2.2	Lag analysis.....	106
4.5.2.3	Price movement analysis	106
4.6	Formulation and formalisation of the trading strategy	107
4.7	Model and trading strategy testing	109
4.8	Conclusion.....	111
Chapter 5	Data collection and preparation	112
5.1	Introduction	112
5.2	Data collection.....	113
5.2.1	Stock selection and historical stock prices	113
5.2.1.1	Hammerson PLC	115
5.2.1.2	Redefine Properties	116
5.2.1.3	FirstRand Limited.....	118
5.2.1.4	Equites Property Fund	119
5.2.1.5	Finbond Group Limited.....	120
5.2.1.6	Growthpoint Properties	122
5.2.1.7	Standard Bank Group	123
5.2.2	Social media data	125

5.3	Data pre-processing.....	131
5.3.1	Stock price data.....	131
5.3.2	Social media data.....	133
5.3.2.1	Data preparation.....	133
5.3.2.2	Feature extraction.....	134
5.3.2.3	Label assignment.....	135
5.4	Conclusion.....	136
Chapter 6	Model development and evaluation	137
6.1	Introduction	137
6.2	Model selection	138
6.2.1	Threshold analysis.....	138
6.2.1.1	Pre-processing techniques.....	139
6.2.1.2	Class distribution.....	141
6.2.2	Selection of models for further analysis	142
6.3	Model development.....	143
6.3.1	Impact on sentiment distribution	143
6.3.2	Hyperparameter optimisation.....	144
6.3.2.1	Logistic regression.....	145
6.3.2.2	Random forests.....	151
6.3.2.3	Long Short-Term Memory	159
6.4	Final model	167
6.5	Conclusion.....	169
Chapter 7	Dynamic trading strategy for the South African stock market.....	170
7.1	Introduction	171
7.2	Keyword extraction and strategic mapping.....	171
7.2.1	Extraction of keywords by sentiment category	171
7.2.2	Mapping to strategic frameworks	174
7.2.2.1	PESTEL analysis	174
7.2.2.2	Porter's Five Forces.....	176
7.3	Stock-specific relationship analysis results.....	178
7.3.1	Hammerson PLC	178
7.3.2	Redefine Properties.....	179
7.3.3	FirstRand Limited	180

7.3.4	Equites Property Fund	181
7.3.5	Finbond Group Limited	182
7.3.6	Growthpoint Properties	183
7.3.7	Standard Bank Group	184
7.4	Emotional valence and cluster analysis	185
7.4.1	Emotional associations with sentiment	186
7.4.2	K-means clustering of emotional signals.....	189
7.5	Thresholds for sentiment intensity classification.....	194
7.6	Trading signals.....	195
7.7	The dynamic trading strategy.....	196
7.8	Testing and evaluation.....	200
7.8.1	Approach	200
7.8.2	Results	201
7.9	Conclusion.....	202
Chapter 8	Conclusions and recommendations	203
8.1	Introduction	203
8.2	Synopsis of the study	203
8.3	Contributions of the study.....	206
8.4	Limitations and future research	208
8.5	Conclusion.....	210
BIBLIOGRAPHY.....		211
APPENDIX A		253

LIST OF FIGURES

Figure 1.1:	Visualisation of the modified CRISP-DM methodology	23
Figure 1.2:	Structure of the study	24
Figure 3.1:	Visualisation of sentiment analysis approaches	58
Figure 3.2:	The functions of the four common paradigms of machine learning (Camps <i>et al.</i> , 2022).....	62
Figure 3.3:	Comparison of the output of classification vs regression analyses (Belkacem, 2021).....	64
Figure 3.4:	Visualisation of the decision tree algorithm (Abdulazeez & Jijo, 2021)	66
Figure 3.5:	Visualisation of the k-nearest neighbour algorithm (Bansal <i>et al.</i> , 2022).....	68
Figure 3.6:	Visualisation of the support vector machine algorithm (Ghosh <i>et al.</i> , 2023).....	71
Figure 3.7:	Visualisation of the sigmoid function (Deshpande <i>et al.</i> , 2023).....	74
Figure 3.8:	The random forest algorithm (Deshpande <i>et al.</i> , 2023)	77
Figure 3.9:	A single hidden layer shallow neural network vs a <i>j</i> -hidden layer deep neural network (Dong <i>et al.</i> , 2019)	78
Figure 3.10:	An example of a deep neural network	79
Figure 3.11:	Architecture of a CNN (Phung & Rhee, 2019)	82
Figure 3.12:	The function of a single computing unit in a convolutional layer (Cui, 2018).....	83
Figure 3.13:	The max vs average pooling approaches (Branko <i>et al.</i> , 2022).....	84
Figure 3.14:	Example of a fully connected layer linked to an output layer for the identification of a car (Taye, 2023)	85
Figure 3.15:	Architecture of an RNN (Wu, 2022).....	87
Figure 3.16:	The operation of an LSTM operation (Rahuljha, 2020).....	88
Figure 4.1:	Training, validation and testing dataset split.....	99
Figure 4.2:	Visualisation of the k-means clustering algorithm (Abbas <i>et al.</i> , 2018)	108
Figure 4.3:	Example of the elbow method (Fan <i>et al.</i> , 2016)	109
Figure 5.1:	2023 Stock performance and trading volume trends for Hammerson PLC .	115
Figure 5.2:	2023 Stock performance and trading volume trends for Redefine Properties	117
Figure 5.3:	2023 Stock performance and trading volume trends for FirstRand Limited	118
Figure 5.4:	2023 Stock performance and trading volume trends for the Equites Property Fund	119
Figure 5.5:	2023 Stock performance and trading volume trends for the Finbond Group Limited	121

Figure 5.6:	2023 Stock performance and trading volume trends for the Growthpoint Properties	122
Figure 5.7:	2023 Stock performance and trading volume trends for the Standard Bank Group.....	124
Figure 6.1:	Visual representation of the sentiment thresholds considered	139
Figure 6.2:	Confusion matrix for the initial logistic regression	146
Figure 6.3:	AUC-ROC curve for the initial logistic regression model.....	147
Figure 6.4:	The bias-variance trade-off for validation loss of the initial logistic regression model	148
Figure 6.5:	The bias-variance trade-off for validation loss of the revised logistic regression model	149
Figure 6.6:	Confusion matrix for the revised logistic regression model	150
Figure 6.7:	AUC-ROC curve for the revised logistic regression model	151
Figure 6.8:	Confusion matrix for the initial random forests model	152
Figure 6.9:	AUC-ROC curve for the initial random forests model	154
Figure 6.10:	Bias-variance trade-off for validation loss of the initial random forests model.....	154
Figure 6.11:	The bias-variance trade-off for validation loss of the revised random forests model	156
Figure 6.12:	Confusion matrix for the revised random forests model.....	157
Figure 6.13:	AUC-ROC curve for the revised random forests model	158
Figure 6.14:	Visualisation of the initial LSTM model.....	159
Figure 6.15:	Confusion matrix for the initial LSTM model	160
Figure 6.16:	AUC-ROC curve for the initial LSTM model	161
Figure 6.17:	Bias-variance trade-off for validation loss of the initial LSTM model	162
Figure 6.18:	Visualisation of the revised LSTM model.....	163
Figure 6.19:	The bias-variance trade-off for validation loss of the revised LSTM model .	164
Figure 6.20:	Confusion matrix for the revised LSTM model.....	165
Figure 6.21:	AUC-ROC curve for the revised LSTM model	166
Figure 6.22:	Example of 10-fold cross-validation.....	168
Figure 6.23:	10-fold cross-validation accuracies of the LSTM model.....	168
Figure 7.1:	Elbow graph for the negative sentiment	189
Figure 7.2:	Elbow graph for the neutral sentiment.....	190
Figure 7.3:	Elbow graph for the positive sentiment.....	190
Figure 7.4:	The clusters of negative sentiment.....	192
Figure 7.5:	The clusters of positive sentiment	193
Figure 7.6:	The final conditions for trading signal generation	196

Figure 7.7:	The dynamic sentiment-directed machine learning trading strategy	198
-------------	--	-----

LIST OF TABLES

Table 1.1:	Secondary objective and corresponding chapter	20
Table 3.1:	Activation functions used in neural networks (Athaiya <i>et al.</i> , 2020; Laxkar & Tomar, 2022)	81
Table 4.1:	Example of a binary confusion matrix (Doulamis <i>et al.</i> , 2021; Kanozia <i>et al.</i> , 2023)	101
Table 4.2:	Expanded multiclass confusion matrix (Afaq <i>et al.</i> , 2020; Doulamis <i>et al.</i> , 2021)	103
Table 4.3:	Correlation interpretation summary (Setiawan & Sina, 2023)	106
Table 5.1:	Example of historical stock data extracted from Yahoo Finance Hub	114
Table 5.2:	Summary of the liquidity and volatility analysis	114
Table 5.3:	2023 Statistics on Hammerson PLC	116
Table 5.4:	2023 Statistics on Redefine Properties	117
Table 5.5:	2023 Statistics on FirstRand Limited	119
Table 5.6:	2023 Statistics on the Equites Property Fund	120
Table 5.7:	2023 Statistics on the Finbond Group Limited	121
Table 5.8:	2023 Statistics on the Growthpoint Properties	123
Table 5.9:	2023 Statistics on the Standard Bank Group	124
Table 5.10:	Checklist for selecting the appropriate data source	125
Table 5.11:	RSS feed links of South African newspapers	127
Table 5.12:	Sample data extracted from RSS feeds	129
Table 5.13:	Sample: data extracted using the web mining approach	131
Table 5.14:	Example of transformed stock price data	132
Table 5.15:	Results of pre-processing	134
Table 5.16:	Vectorisation of the sample text strings	134
Table 5.17:	Threshold configurations for sentiment label assignment scenarios	135
Table 6.1:	Prediction accuracy (%)	140
Table 6.2:	Class distributions across thresholding scenarios	141
Table 6.3:	Classification report for the initial logistic regression model	147
Table 6.4:	Upper and lower bounds of hyperparameters in the revised logistic regression model	149
Table 6.5:	Classification report for the revised logistic regression model	151
Table 6.6:	Classification report for the initial random forests model	153

Table 6.7:	Upper and lower bounds of hyperparameters in the revised random forests model	156
Table 6.8:	Classification report for the revised random forests model	158
Table 6.9:	Classification report for the initial LSTM model.....	161
Table 6.10:	Upper and lower bounds of hyperparameters in the revised LSTM model..	163
Table 6.11:	Classification report for the revised LSTM model	166
Table 7.1:	Keywords associated with a negative sentiment	172
Table 7.2:	Keywords associated with neutral sentiment.....	173
Table 7.3:	Keywords associated with positive sentiment.....	173
Table 7.4:	Keywords mapped to the PESTEL framework.....	175
Table 7.5:	Keywords mapped to Porter's Five Forces	177
Table 7.6:	Relationship analysis for Hammerson PLC stock	179
Table 7.7:	Relationship analysis for Redefine Properties	180
Table 7.8:	Relationship analysis for FirstRand Limited.....	181
Table 7.9:	Relationship analysis for Equites Property Fund	182
Table 7.10:	Relationship analysis for Finbond Group Limited	183
Table 7.11:	Relationship analysis for Growthpoint Properties	184
Table 7.12:	Relationship analysis for Standard Bank Group	185
Table 7.13:	Summary of characteristics of emotion analysis.....	186
Table 7.14:	Descriptive statistics of emotions across sentiment classes	187
Table 7.15:	Calculated characteristics of negative and positive sentiments	195
Table 7.16:	Performance of the dynamic trading strategy	201

LIST OF ALGORITHMS

Algorithm 3.1:	Forward propagation in a neural network	79
Algorithm 3.2:	Backpropagation in a neural network.....	80
Algorithm 5.1:	Data collection for historical RSS feeds	126
Algorithm 5.2:	Web mining historical news	129
Algorithm 7.1:	Dynamic trading strategy for the South African stock market	199

LIST OF ACRONYMS

AI	Artificial Intelligence
API	Application programming interface
AUC	Area under the curve
AUC-ROC	Area Under the Receiver Operating Characteristic

BERT	Bidirectional Encoder Representations from Transformers
CART	Classification and Regression Trees
C4.5	Classification 4.5
CNN	Convolutional neural network
CRISP-DM	Cross Industry Standard Process for Data Mining
DistilBERT	A lighter version of BERT
ESG	Environmental, Social, and Governance
FinBERT	Financial sentiment analysis model based on BERT
FN	False negatives
FP	False positives
FTSE	Financial Times Stock Exchange
GDP	Gross Domestic Product
HTML	HyperText Markup Language
ID3	Iterative Dichotomiser 3
JSE	Johannesburg Stock Exchange
LSTM	Long Short-Term Memory
LSTMGA	LSTM model optimised with a genetic algorithm
MAE	Mean Absolute Error
NASDAQ	National Association of Securities Dealers Automated Quotations
PESTEL	Political, economic, social, technological, environmental, and legal
PPV	Positive predictive value
RMSE	Root Mean Square Error
RNN	Recurrent neural network
ROC	Receiver Operating Characteristic
R^2	Coefficient of Determination
RSS	Really Simple Syndication
SSE	Sum of squared Euclidean distances
TF-IDF	Term Frequency - Inverse Document Frequency
TN	True negatives
TP	True positives
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary for Sentiment Reasoning
XGBoost	Extreme Gradient Boosting
XML	Extensible Markup Language

CHAPTER 1

Introduction and problem contextualisation

1.1	Introduction	15
1.2	Problem statement	16
1.3	Secondary research questions.....	18
1.4	Contributions.....	20
1.5	Scope and limitations	21
1.6	Structure of the thesis	24
1.6.1	Phase I – Introduction.....	24
1.6.2	Phase II – Business understanding.....	24
1.6.3	Phase III – Data understanding and preparation.....	25
1.6.4	Phase IV – Model development and evaluation	25
1.6.5	Phase V – Trading strategy formulation and deployment.....	25
1.6.6	Phase VI – Conclusion	25
1.7	Conclusion.....	26

1.1 Introduction

The purpose of this chapter is to establish the background and motivation for developing a dynamic trading strategy that applies a machine learning framework to model social media sentiment for stock market trading in South Africa. The chapter begins with an introduction to the problem statement by highlighting some key challenges in sentiment analysis for stock market trading, specifically in the context of the South African market.

Despite the growing importance of sentiment analysis and machine learning in global stock market prediction, limited research focusing specifically on these techniques for the South African environment exists. The need to develop such a trading strategy originated due to the unique structural, behavioural, and sentiment-driven characteristics of the Johannesburg Stock Exchange (JSE), which markedly differed from those of developed markets (Nyakurukwa & Seetharam, 2023). The characteristics include lower liquidity across many counters, concentrated sectoral exposure, and heightened sensitivity to local economic and political events. Moreover, the scarcity of sentiment-driven machine learning strategies in an emerging-market context

indicated a clear gap in the literature (Ferreira-Schenk *et al.*, 2025). Designing a market-specific strategy, therefore, would ensure alignment with local information flows, investor behaviour, and practical trading conditions relevant to South African market participants. This gap was especially evident in the absence of comprehensive publications on the integration of sentiment analysis, deep learning applications, and trading strategies tailored to the South African market.

Some background on the problem is provided in Section 1.2, which concludes with the research question. The secondary research questions are listed in Section 1.3, followed by a layout of the contributions of the study in Section 1.4 and the scope and limitations in Section 1.5. A complete structure of the thesis is presented in Section 1.6.

1.2 Problem statement

The stock trading process used to be entirely manual and relied heavily on intuition (Agrawal *et al.*, 2013). As trading increased in popularity and competition intensified, investors started searching for tools and methods to maximise profits while minimising risk (Pawaskar, 2022). Along with the fourth industrial revolution, demand for digital transformation with access to existing and new markets became paramount (Kern & Sullivan, 2021). This opened the door for online trading, which has become quite common. Intuition has been transformed into more statistically based methods that incorporate fundamental (long-term) and technical (short-term) analyses (Gupta *et al.*, 2021). However, these methods were based on the assumption that investors behaved rationally, which was not true (Chaudhary & Vaid, 2022; De Bondt, 2020). Moreover, what frightened future and current investors the most about stock markets was the extreme implied volatility that caused variability in stock prices over time (Mamtha & Srinivasan, 2016). Historical volatility can be mathematically calculated as the standard deviation of a stock's average return (Gong *et al.*, 2022). This indicated how closely a stock's price fluctuated around the mean or moving average (Rathburn *et al.*, 2022). Unfortunately, future volatility could not be predicted with complete accuracy, as it was an estimated probability of how stock prices might move (He & Qu, 2022). Combining this with turbulent market dynamics made accurately predicting future stock prices quite challenging (Pawaskar, 2022).

Singpurwala (2021) stated that stock trading was related to human psychology, with an 81% correlation between market sentiment and Bitcoin prices, a decentralised digital currency. The consequent inclusion of market emotionalism, or the influence of emotion on financial markets, may lead to a more comprehensive view of market fluctuations and reduce human confusion (Othalasseril & Shaikh, 2021). The latter was supported by Maqsood *et al.* (2022:21), who stated that "*emotion and the sentiment expressed by the public influence the stock exchanges and the buying patterns of investors, as well as business decisions*". There was a definite and relevant

correlation between stock prices and human emotion (Mendoza-Urdiales *et al.*, 2022). The reason for this was that anxious and distressed traders were more likely to start saving money, which could lead to a decline in the stock market. However, the opposite was also true; happy and content traders were more likely to feel encouraged to purchase stocks, which would cause the stock market to rise because supply was low and demand was high.

A popular text classification tool that has been used to incorporate the quantified version of human psychology, which is the process of numerically measuring human psychological factors, into trading is a form of text mining and analysis, named sentiment analysis (Othalasseril & Shaikh, 2021). The technique has become essential for extracting insights from social networks (Iglesias & Moreno, 2019). Sentiment analysis is the process of identifying, interpreting, and assessing emotions and satisfaction levels in specific contexts using software (Chen & Yao, 2021). There are four main types of sentiment analysis techniques distinguished by the output or purpose of the analysis. Fine-grained methods determine the degree of positivity or negativity in an opinion, aspect-based approaches identify people's opinions on specific elements, intent-based analyses detect the purpose of a specific statement, and emotion detection captures underlying feelings like fear or excitement (Bakal & Shah, 2022; Chua *et al.*, 2022). The latter generally applies to a fundamental analysis of social media data (Gupta *et al.*, 2021).

Sentiment analysis is typically driven by natural language processing (NLP) in combination with machine learning algorithms to detect the emotion or sentiment voiced by someone about a certain topic (Gupta *et al.*, 2023). As computers were incapable of understanding text, NLP was the tool that enabled machines to digest and analyse language as humans do by transforming significant words into numbers (Hlomani *et al.*, 2016). Various machine learning models could be used or even combined to perform sentiment analysis, with the Long Short-Term Memory (LSTM) model proving to be very popular and successful (Fuller, 2022; Gupta *et al.*, 2021). A *no-forget* feature allowed it to retain past stock prices, which was essential for predicting future stock trends (Othalasseril & Shaikh, 2021). Additionally, it also considered the context of a sentence rather than a single word or part of a sentence (Fuller, 2022). Furthermore, social media commentary or alternative data was becoming more popular in the fields of investment management and algorithmic trading (Hansen & Borch, 2022).

South Africa was classified as a developing country, meaning it faced economic challenges such as a 3.2% post-COVID-19 economic contraction (Chaminuka *et al.*, 2022; Nortje, 2017). These conditions, along with unique political, social, and regulatory dynamics, rendered trading strategies from developed markets like the United States of America less effective due to differences in spending habits, macroeconomic variables, and market sentiment drivers. The advantages of a successful trading strategy have been proven in the real estate, healthcare, food,

and software sectors (Campbell *et al.*, 2019; Mazur *et al.*, 2021; Poterba, 2000). Additionally, in the South African context, sentiment-driven approaches may hold particular significance due to the country's unique political, economic, and social landscape (Ferreira-Schenk *et al.*, 2025; Muguto *et al.*, 2019; Nyakurukwa & Seetharam, 2023).

The goal of this study is to develop a trading strategy for selected South African stocks by implementing a machine learning approach to perform sentiment analysis on social media and model stock movement, accordingly. Such a strategy would make valuable contributions by enhancing personal wealth, which in turn would support company growth and attract international investors. This translates to the following research question: "To what extent can a trading strategy be based on social media sentiment in the South African stock market?"

1.3 Secondary research objectives

The main aim of the study is to combine the use of machine learning techniques and sentiment analysis on social media data to develop a dynamic trading strategy for selected stocks on the South African stock market. Towards achieving this, the secondary objectives that need to be met are as follows:

- *Conduct a detailed literature study on the stock market.*
An overview of the structure and functioning of stock exchanges, with emphasis on the JSE as the focal point of the South African financial market, is presented in Chapter 2. This includes a discussion on how macroeconomic, microeconomic, firm-specific, and global factors influence stock prices and discusses the role of investor perception and sentiment in shaping market behaviour. The purpose of this chapter is to establish the theoretical foundation linking sentiment to supply-demand dynamics in financial markets.
- *Conduct a comprehensive literature study on sentiment analysis, data extraction from social media platforms, and machine learning in stock market trading.*
In Chapter 3, sentiment analysis in financial contexts is reviewed, including data extraction from platforms such as Twitter/X, Reddit, and online news sources. It covers pre-processing, feature extraction, sentiment labelling, and ethical considerations. Traditional machine learning models like decision trees, logistic regression and support vector machines and advanced deep learning models including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are introduced. The chapter concludes by identifying research gaps relevant to emerging markets such as South Africa.
- *Establish a hybrid sentiment analysis and machine learning research methodology.*
Directed by the literature in Chapters 2 and 3, the hybrid research methodology for this study is described in Chapter 4.

- *Select stocks eligible for use in this study.*

Stock selection is based on liquidity, volatility, and the availability of sufficient historical data. The criteria and approach, guided by the literature, are described in Chapter 5.

- *Identify reliable social media platforms suitable for ethical data collection.*

Chapter 5 continues with an in-depth discussion on the principles and approach for selecting social media and news-based sources. Platforms were evaluated for relevance, accessibility, temporal consistency, data quality, and ethical compliance.

- *Collect and process the data along with the selected stock price history.*

The data collection process is further covered in Chapter 5, which includes the parallel collection of sentiment-bearing text data and historical stock price data. Both datasets were synchronised to correspond to the same time period. This is followed by the pre-processing techniques evaluated and utilised for data cleaning and preparation.

- *Perform sentiment analysis on the collected data.*

Towards creating the systematic outline for the trading strategy, the sentiment analysis process is discussed in Chapter 6. Data-directed sentiment labels are established.

- *Configure and evaluate machine learning models for suitability in sentiment classification and prediction.*

Chapter 6 includes the approach for configuring and evaluating the machine learning models best suited to the collected data. The hyperparameter optimisation and performance evaluation processes are accordingly discussed.

- *Extract and map keywords to strategic frameworks.*

Upon establishing the ideal combination of sentiment analysis and machine learning models, keyword extraction and mapping to macroeconomic and microeconomic frameworks are discussed in Chapter 7. This step serves to verify that the sentiment-derived keywords correspond to strategic powers relevant to each stock.

- *Verify the model outputs using appropriate statistical techniques.*

The selected best-performing developed model is verified using correlation analysis, lag analysis, and price-position analysis, as also described in Chapter 7. The process includes evaluating the relationships between sentiment and stock price behaviour. These analyses validate the robustness and explanatory power of the predictive models.

- *Formulate, evaluate and test the trading strategy for the selected stock(s) in the South African stock market.*

Chapter 7 culminates in a detailed description of the sentiment-driven dynamic trading strategy, resulting from the extensive analyses. The strategy is presented as a Unified Modelling Language (UML) diagram and formalised in an algorithm.

Table 1.1 provides a summary of the research objectives and the specific chapters in which each objective is addressed. This summary provides the reader with a clear roadmap of how the study is structured and ensures that the progression from the problem statement to the conclusions is presented logically.

Table 1.1: Secondary objective and corresponding chapter

Summary of study contributions	
Objective	Chapter
Conduct a detailed literature study on the stock market.	2
Conduct a comprehensive literature study on sentiment analysis, data extraction from social media platforms and machine learning in stock market trading.	2, 3
Establish a hybrid sentiment analysis and machine learning research methodology.	4
Select stocks eligible for use in this study.	5
Identify reliable social media platforms suitable for ethical data collection.	5
Collect and process the data along with the selected stock price history.	5
Perform sentiment analysis on the collected data.	5
Configure and evaluate machine learning models for suitability in stock price prediction.	6
Extract and map keywords to strategic frameworks	7
Verify the model outputs using appropriate statistical techniques.	7
Formulate, evaluate and test the trading strategy for the selected stock(s) in the South African stock market.	7

The eleven objectives presented in Table 1.1 collectively forms the framework that guided the research process. Each objective was designed to address a specific component of the research problem, leading to a structured pathway to achieving the aim of this study. By following these objectives, the primary research question will be systematically addressed.

1.4 Contributions

This study makes several contributions to sentiment analysis and machine learning in stock market trading, with a specific focus on the South African context. The contributions are methodological, empirical, and practical, and they establish a structured framework for transforming sentiment and emotion signals into actionable trading decisions.

- A sentiment-driven trading strategy tailored to the South African stock market is developed. By translating sentiment and emotion scores into buy, hold, and sell signals, the work bridges the gap between predictive modelling and practical decision-making and represents one of the first strategies designed and tested specifically for this market.

- An inverse keyword extraction approach is adopted by correlating stock price movements with news headlines to derive the most relevant sentiment-bearing keywords. This allows for the identification of clear associations between sentiments and emotions, both negative and positive, which are validated through statistical thresholding methods.
- A hybrid sentiment analysis methodology is introduced, integrating lexicon-based tools with machine learning classifiers, including logistic regression, random forests, and LSTM networks. This combined approach demonstrates how interpretability and predictive power can be achieved simultaneously.
- The contribution relating to data collection and pre-processing practices includes how RSS feeds and web mining can be combined to overcome access, ethical, and completeness constraints. This dual-source dataset supports both model training and correlation-based financial analyses.
- Empirical evidence of the predictive relationship between sentiment and stock price movements in South Africa is presented. Through correlation, lag, and price-position analyses, the study shows how sentiment influences revenue at different intervals, particularly in the post-COVID-19 period.
- A comprehensive methodological framework that integrates sentiment classification, emotion detection, cluster-based thresholding, and trading rule generation is established. This framework offers a replicable, data-driven foundation for sentiment-informed trading in emerging markets.

Collectively, these contributions advance the use of sentiment and machine learning in financial modelling, while addressing the lack of localised, empirically validated strategies for the South African stock market. This is a valuable contribution to the international field of sentiment-driven machine learning strategies in an emerging-market context.

1.5 Scope and limitations

The work integrates sentiment analysis with machine learning models to generate structured buy, hold, and sell signals based on sentiment and emotional intensity. The scope is defined as follows:

- *Market focus:* The analysis is limited to the JSE, using a selected group of stocks chosen for their liquidity, volatility, and relevance to sentiment-based modelling.
- *Data:* Sentiment was derived from South African online news articles collected through RSS feeds and web mining. This approach ensured access to high-quality, structured textual data suitable for model development and financial analysis.
- *Methodology:* The study employs machine learning techniques that include lexicon-based and supervised models for sentiment classification, emotion analysis, and trading signal generation.

- *Trading strategy*: A sentiment- and emotion-based strategy is developed at the monthly time scale, enabling the transformation of linguistic signals into actionable trading decisions.

Despite its contributions, several limitations must be acknowledged. They include:

- *Data coverage constraints*: RSS feeds contained gaps across certain months, which restricted the completeness of temporal analyses. Accordingly, the data is collected primarily from newspaper-based platforms. While formal media provides structured information, it may not fully capture real-time investor sentiment found on social media.
- *Stock selection*: Testing is performed on a small group of highly liquid stocks. Although they are selected based on criteria listed in established research, this limits the generalisability of the findings to the broader JSE or other asset classes.
- *Temporal aggregation*: Trading signals are generated at a lagged monthly level. This approach is a proof of concept, but it limits the strategy's responsiveness to rapidly changing market conditions.
- *Risk-adjustment limitations*: The sentiment thresholds used for decision-making do not incorporate an individualised risk-adjusted framework. As a result, individual risk preferences and mitigation strategies during adverse market conditions are not modelled.

While these limitations guide the interpretation of the findings, they also highlight clear opportunities for future research in sentiment-based trading within emerging markets. The methodology of the study is grounded in a modified version of the Cross Industry Standard Process for Data Mining (CRISP-DM) to support the formulation of the trading strategy (Dumas *et al.*, 2022). This process includes the following phases (Ferreira *et al.*, 2023):

- *Business understanding*: The focus is on understanding the project's goals and requirements from a business perspective to translate these into a data mining problem and create a project plan.
- *Data understanding and preparation*: Data understanding involves gathering and familiarising with the data to gain initial insights and form hypotheses. However, data preparation focuses on creating a dataset for the modelling phase. The modelling phase involves data selection, pre-processing, transformation, and cleaning, which may be repeated without a fixed sequence.
- *Modelling and evaluation*: Modelling involves comparing techniques and estimating parameters to solve the problem. During this phase, issues with the dataset may be identified, prompting a return to previous phases for data adjustments. The evaluation focuses on validating the chosen model while reviewing previous steps to ensure the model aligns with business goals and predefined metrics.

- *Trading strategy formulation*: The formulation focuses on developing a comprehensive strategy based on insights gained from the data analysis and modelling phases. This step involves selecting suitable trading signals to inform decision-making during the trading process. The formulated strategy is designed to meet the primary objective while optimising performance based on the analysis.
- *Deployment*: Deployment involves packaging the model in a way that enables its practical use. It may involve delivering a solution ranging from a simple report to a dashboard to implementing a repeatable data mining process across the organisation.

The modified CRISP-DM approach is presented in Figure 1.1.

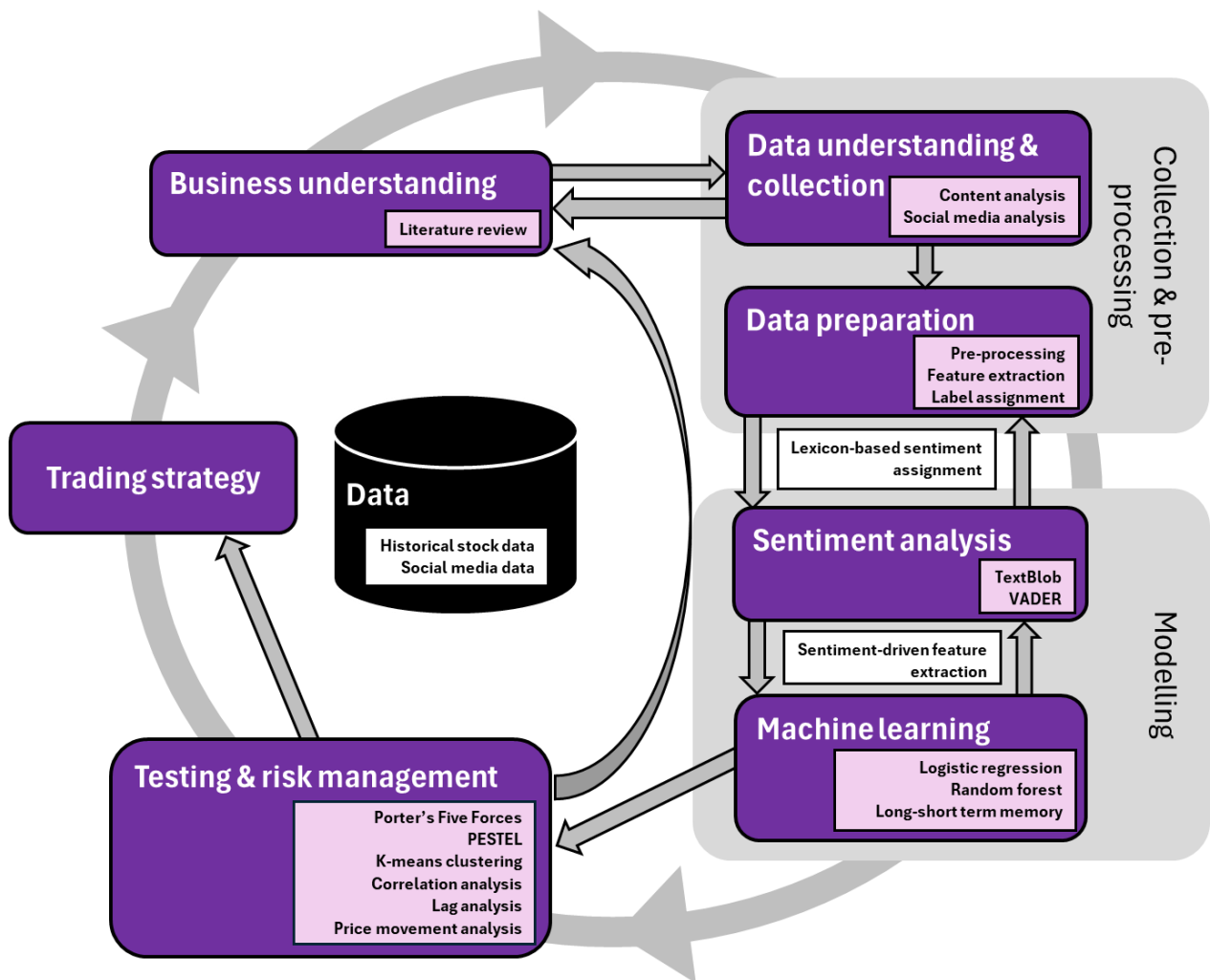


Figure 1.1: Visualisation of the modified CRISP-DM methodology

The research commences by understanding the business context, which in turn sets the stage for examining the stock market, sentiment analysis, and machine learning. This guides *data understanding and preparation*, where raw data will be collected, investigated, cleaned, and processed. The next step involves *model development and evaluation*, where various machine learning models will be built and assessed based on their performance in predicting sentiment.

From model development, significant keywords for each sentiment class are determined and evaluated. Finally, the *trading strategy formulation and deployment* phase integrates the evaluated models into actionable trading signals for implementation in the real-world environment. The approach allows for iteration where previous steps can be revisited to make adjustments, if required.

1.6 Structure of the thesis

The study is structured into nine chapters that are organised according to the modified version of the CRISP-DM. This is illustrated in Figure 1.1.

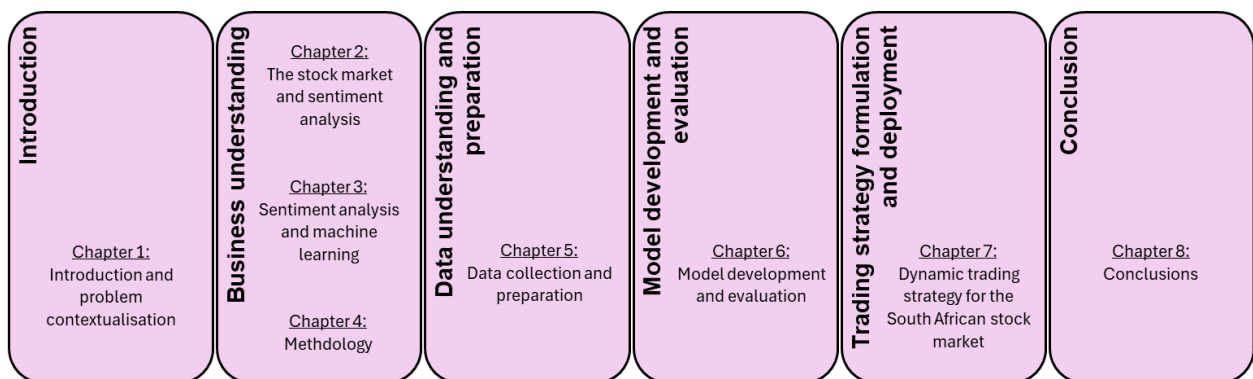


Figure 1.2: Structure of the study

Each phase builds on the preceding one to ensure methodological coherence and logical progression. As discussed in the preceding subsections, the framework transitions from problem contextualisation and theoretical grounding toward data preparation, model development, and ultimately, the formulation and deployment of a dynamic trading strategy tailored to the South African stock market. The final phase concludes the research by consolidating findings and outlining the key contributions of the study.

1.6.1 Phase I – Introduction

The introduction phase encompasses the problem statement, goals and objectives, scope and limitations, and concludes with the structure of the thesis. The introduction is made in Chapter 1.

1.6.2 Phase II – Business understanding

Phase two of the study involves a review of past research, detailed in Chapter 2, 3 and 4. Chapter 2 focuses on the structure and functions of the stock market, stock price determination, and factors that influence stock prices. It also examines sentiment analysis, previous research in sentiment analysis, and key gaps in the literature. Chapter 3 introduces machine learning, defining its purpose and approaches while outlining the methodology. Additionally, it explores the

application of machine learning in sentiment analysis. This phase establishes the foundation for understanding stock market dynamics and sentiment analysis, shaping the methodological framework of the study that is outlined in Chapter 4.

1.6.3 Phase III – Data understanding and preparation

The third phase emphasises the collection and preparation of data for the subsequent analysis. Chapter 5 defines the criteria for social media data collection, detailing the selection of appropriate data sources and outlining the necessary pre-processing techniques to ensure the data's quality. Key activities in this phase include data selection, pre-processing, transformation, and cleaning, all of which ensure that the dataset is both relevant and ready for the modelling phase.

1.6.4 Phase IV – Model development and evaluation

For the fourth phase, the focus is on developing and evaluating machine learning models designed for sentiment analysis and stock market predictions. Chapter 6 discusses the various machine learning models that will be evaluated, describing the training process, evaluation methods, and hyperparameter optimisation. In Chapter 7, analytical methods for relationship analysis are explored, along with the results from sector-specific models. Evaluating the performance of these models is crucial to ensuring they meet the study objectives and are capable of delivering accurate predictions.

1.6.5 Phase V – Formulation and formalisation of the trading strategy

During phase five, a trading strategy is crafted based on the insights derived from the previous data analysis and model evaluations. Chapter 7 continues by examining the identification of significant keywords, emotional associations with each sentiment, and the definition of trading signals, alongside the establishment of thresholds for emotional signals. The strategy undergoes testing and validation to ensure it performs effectively under real-world conditions, providing a structured approach to informed trading decisions.

1.6.6 Phase VI – Conclusion

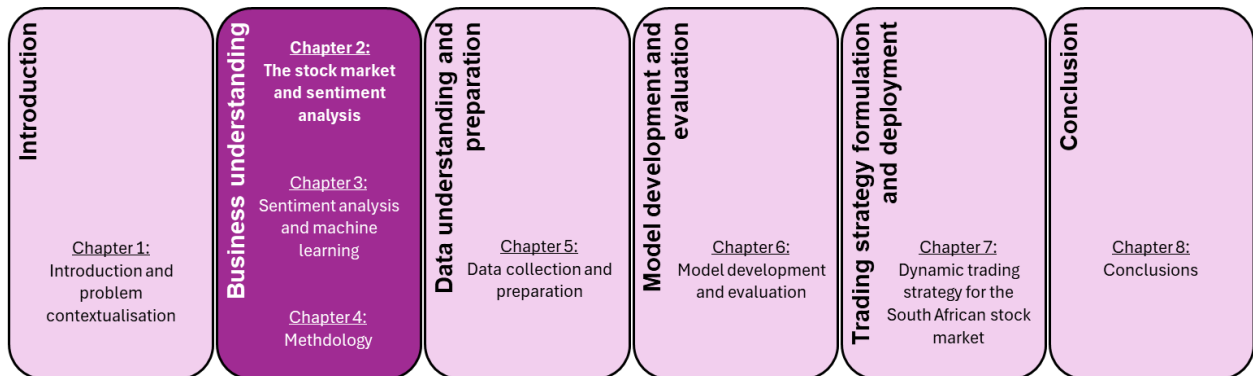
In the final phase, the trading strategy is deployed, and the study's conclusions are drawn. Chapter 8 reviews the key findings, discusses the contributions of the research, and offers recommendations for future inquiries. The deployment phase focuses on packaging the strategy for practical use. The conclusion synthesises the research, reflecting on its outcomes and implications for the field.

1.7 Conclusion

The purpose of this chapter was to introduce the core of the study by outlining the problem statement, goal and objectives. Next, the scope and limitations of the research were clarified. The chapter concluded with an overview of the thesis structure, summarising the main themes of each phase and briefly describing the topics to be covered in each chapter.

CHAPTER 2

The stock market and sentiment analysis



2.1	Introduction	27
2.2	The stock market.....	28
2.2.1	Overview of stock exchanges	29
2.2.2	The Johannesburg Stock Exchange	30
2.3	Stock price determination.....	33
2.4	Factors influencing stock prices.....	34
2.4.1	Macroeconomic factors.....	34
2.4.2	Microeconomic factors and firm-specific factors.....	35
2.4.3	Global events and market shocks	35
2.5	Sentiment analysis	36
2.5.1	Introduction.....	37
2.5.2	Sentiment analysis in the stock market.....	40
2.5.3	Sentiment analysis in the South African context	44
2.6	Machine learning	48
2.6.1	Introduction.....	48
2.6.2	Recent research in machine learning.....	49
2.7	Conclusion.....	52

2.1 Introduction

In order to fully understand how to approach stock price prediction for the South African market using sentiment analysis and machine learning, these categories all need to be examined further.

This chapter represents the second phase in the presentation of the study. It covers the background of the stock market and the role of sentiment analysis in making informed trading decisions, as well as its use in South Africa, and the use of machine-learning to enhance the accuracy, scalability and efficiency of sentiment analysis.

A comprehensive overview of stock exchanges is presented in Section 2.2, which focuses on the Johannesburg Stock Exchange (JSE), as well as the key factors influencing stock prices in Section 2.3 and 2.4. Understanding these market dynamics is essential for interpreting how sentiment can influence stock market behaviour. Section 2.5 presents a discussion on sentiment analysis, its relevance to stock market predictions, and its application in assessing investor sentiment. Gaps in the current literature are identified in Section 2.6, particularly in the context of South African markets.

2.2 The stock market

Defining a stock market means considering three dependent/relevant sections, namely stock exchanges, listed companies, and stock prices (Prats & Sandoval, 2020). The stock market represents stock exchanges that are diverse marketplaces formed from a network of economic transactions where financial securities are bought and sold between participants at an agreed price (Aničić *et al.*, 2018). Listed companies can either be private or public, subdividing stock exchanges into two types of markets: a primary market and a secondary market (Andriansyah, 2017). A primary market is one where private companies offer their shares to the public and get listed on the stock exchange, allowing investors to purchase their stocks (Kumar & Nidugala, 2020). Moreover, the transaction chain of the primary market consists of four stages: fund-raising, asset investment, asset monitoring, and continuation (Chen *et al.*, 2022). Fund-raising involves securing capital from investors or institutions to support a defined investment strategy. Once capital is raised, the asset investment phase entails allocating these funds to selected assets to achieve targeted returns. Following investment, asset monitoring is conducted to evaluate the performance, risk exposure, and compliance of the assets to ensure they remain aligned with strategic goals. The final stage, continuation, involves making informed decisions on whether to maintain, adjust, or exit the investment based on performance assessments and prevailing market conditions. The process by which companies list their shares on a stock exchange is an Initial Public Offering (Christiansen & Koldertsova, 2009). In the secondary market, existing stocks are bought and sold among investors (Kumar & Nidugala, 2020). This section covers an overview of stock exchanges and the JSE, which is central to South Africa's financial markets.

2.2.1 Overview of stock exchanges

The primary function of a stock exchange is to trace the supply, demand, and price movements of stocks of listed companies, and it serves as the platform where both short-term trading and long-term investing take place (Masoud, 2013). Moreover, it is a regulated marketplace where listed securities of corporations, governments and other entities are bought and sold by investors and speculators through authorised members, operating under established rules and governance by a board of directors (Abraham & Kannappan, 2018).

Stock market trading has evolved significantly since its early forms in 12th-century France, where brokers known as *courtiers de change* were responsible for managing agricultural debts (Bhatti *et al.*, 2023). However, the first formal stock market emerged in 1460 in Antwerp, Belgium, followed by the Amsterdam Stock Exchange in 1602, which marked the beginning of modern stock trading (Escobar *et al.*, 2013; Maghanga & Quisenberry, 2015). The rise of the Dutch East India Company, which initiated the first stock exchange, further cemented Amsterdam's role in stock market history (Gelderblom & Jonker, 2004; Vergne, 2008). The London Stock Exchange developed in the 1700s, while the New York Stock Exchange was established in 1817 (Nyasha & Odhiambo, 2013; Stringham, 2002). Additionally, the Industrial Revolution spurred the demand for stock markets, with exchanges becoming essential for raising capital and economic growth (Campbell & Rogers, 2017; Michie, 1986). Other key exchanges followed, such as the Toronto Stock Exchange in 1852 and the JSE in 1887 (De Beer *et al.*, 2015). Technological advancements in the 20th century, including the founding of the National Association of Securities Dealers Automated Quotations (NASDAQ) in the United States of America in 1971, transformed trading practices (Ocran, 2010). The collapse of the stock market, known as the Great Crash of 1929, spurred by speculation and economic instability, led to regulatory changes such as the Securities Exchange Act of 1934 (Benston, 1973; Doty, 1992). Likewise, the crash of 1987 altered market psychology with a shift toward reduced risk aversion until the 'technology bubble' in the year 2000, when investors started pouring money into internet-based companies (Boyer, 2009). Nowadays, remote access, high-frequency trading, and the rise of blockchain and cryptocurrencies are reshaping global markets (Basma, 2024; Chung & Lee, 2016; Guney *et al.*, 2017; Hameed, 2019).

The stock market has evolved from informal, local exchanges to a highly sophisticated global financial system. Stock exchanges have evolved into institutions that facilitate substantial transactions on a daily basis, significantly shaping economies worldwide. With technological advances, the future of stock markets is focused on digital trading and decentralised finance. Amidst this global evolution, regional exchanges like the JSE have played an essential role in bridging local investment opportunities with the broader international financial landscape.

2.2.2 The Johannesburg Stock Exchange

The JSE is the 19th largest stock exchange globally and the largest in Africa by market capitalisation (Becker, 2024; Chung *et al.*, 2013; Holden *et al.*, 2016). This section provides an overview of the JSE by examining its historical development, economic significance, market structure, and key indices. As the largest stock exchange in Africa, the JSE plays a crucial role in South Africa's financial markets, facilitating capital raising, investment, and economic growth (Chung *et al.*, 2013; Lukasiewicz, 2017).

2.2.2.1 Historical development

The JSE was established in 1887 in response to the rapid expansion of the South African mining sector, particularly the discovery of diamonds and gold during the Gold Rush in the Witwatersrand, South Africa (De Beer *et al.*, 2015; Holden *et al.*, 2016). For more than 120 years, the JSE played a crucial role in the development of financial and capital markets in South Africa (Chung *et al.*, 2013). Since then, it has been shaped by significant shifts in domestic policy, global monetary changes, economic fluctuations, and market crises, including apartheid and the Great Depression (De Beer *et al.*, 2015).

Following the introduction of financial market legislation in 1947, the JSE joined the World Federation of Exchanges in 1963 and transitioned to electronic trading in the early 1990s (Mpini & Tchereni, 2020). Before its formalisation, securities were traded informally. However, the JSE was established to provide a structured platform for financing mining operations that played a key role in South Africa's transition to an industrialising, mineral-export-driven economy (De Beer *et al.*, 2015). By the late 20th century, the JSE had, over the decades, expanded beyond mining to include industries such as banking, retail, telecommunications, and manufacturing (Lukasiewicz, 2017). In 1996, the JSE demutualised (Verhoef, 2012) or transitioned from a member-owned to a shareholder-owned institution with improved operational efficiency, and strengthened its competitiveness in global financial markets (Irving, 2005). Additionally, the adoption of the London Stock Exchange electronic trading system in the same year replaced the traditional open-outcry floor and streamlined market operations (Dicle & Levendis, 2013).

The operations at the JSE have consistently been modernised through technological advancements (Ebersohn *et al.*, 2017). Following the adoption of electronic trading, the exchange has further incorporated high-frequency trading, which is a subset of algorithmic trading, as well as real-time data analytics (Courdent & McClelland, 2022). These advancements have enhanced market efficiency, liquidity, and price discovery while expanding the JSE's digital infrastructure to support growing transaction volumes and investor access (Britten & McKane, 2018; Mhlongo *et al.*, 2024; Muller & Ward, 2013). These developments have strengthened the JSE's position as a

leading financial institution in Africa, aligning its market infrastructure with global standards and enhancing its competitiveness in the international monetary system (Chung *et al.*, 2013; De Beer *et al.*, 2015; Lukasiewicz, 2017).

2.2.2.2 Economic significance

Stock markets and stock exchanges are of crucial importance to a country's economic growth (Masoud, 2013). Financial markets facilitate economic development by serving as intermediaries between savings and investments, especially in countries where access to capital is critical (Dicle & Levendis, 2013). Heertje (2004) explains that the government and companies can raise long-term capital by selling stocks on the primary stock market, which leads to an improved Gross Domestic Product (GDP) because capital enables companies to expand, grow, and stimulate employment (El-Wassal, 2013). Therefore, stock market performance is a significant indicator of economic performance and the cycle that the economy resides in, like a recession or expansion (or *boom*) (Alam *et al.*, 2016).

By enabling capital raising for both public and private sector entities, the JSE plays a fundamental role in South Africa's financial landscape (Heertje, 2004). As Africa's most liquid stock exchange, the JSE attracts substantial foreign investment and serves as a gateway for international capital flows into the region (De Beer *et al.*, 2015; Lukasiewicz, 2017). Through its regulatory framework and listing requirements, the exchange facilitates corporate access to investment capital while enabling investors to diversify their portfolios (Benjelloun, 2017). According to the World Economic Forum's 2014-2015 Global Competitiveness Index Survey, South Africa is ranked first in securities exchange regulations out of 144 countries (Holden *et al.*, 2016). The exchange facilitates bond issuance, which enables government and corporate debt financing (Ngwenya & Simatele, 2020). By improving transparency and liquidity, the JSE helps strengthen financial stability and supports economic growth in South Africa (Benjelloun, 2017; Firer & Neu-Ner, 1997; Heertje, 2004).

2.2.2.3 Market structure and key indices

The JSE consists of multiple market segments that support a wide range of financial instruments and issuers while ensuring access to capital for companies of different sizes and investment needs, namely the main board, Alternative Exchange, bond market and derivatives, and commodity markets (Adelegan, 2009; Merino & Sebastian, 2019; Ojah & Oluoch, 2024). The main board is the primary market for large, well-established companies, giving them access to substantial capital and liquidity while offering investors opportunities to invest in mature and stable enterprises. The Alternative Exchange caters to more minor, high-growth companies seeking to raise capital and expand their market presence. It serves as an entry point for emerging

businesses, enabling progression to the main board upon meeting stricter regulatory standards. The bond market facilitates the issuance and exchange of corporate and government debt securities. It allows public and private sector entities to raise capital efficiently while providing investors with fixed-income opportunities. The derivatives and commodity markets encompass a range of financial instruments, which include equity and interest rate derivatives, as well as agricultural commodity contracts. These markets help investors manage risk, hedge against price fluctuations and pursue speculative opportunities (Merino & Sebastian, 2019). With a structured yet diverse market framework, the JSE supports various financial activities and remains a comprehensive platform for trading equities, fixed-income instruments, and derivatives (Adelegan, 2009; Chitundu, 2022). This multi-segment structure strengthens market depth, liquidity, and accessibility, thereby solidifying the JSE's position as a leading financial institution in Africa.

A range of benchmark indices maintained by the JSE offers valuable insights into market performance and investor sentiment (Sgammini, 2023). The Financial Times Stock Exchange (FTSE)/JSE Africa All Share Index reflects the overall performance of all companies listed on the exchange, providing a broad view of the market's health. Tracking the performance of the 40 largest companies by market capitalisation, the JSE Top 40 Index focuses on the most prominent players in the South African economy. The FTSE/JSE Responsible Investment Index comprises companies that meet specific environmental, social, and governance criteria while catering to investors who prioritise sustainable and socially responsible investment practices. These indices are crucial performance indicators for investors, fund managers and policymakers, providing them with the necessary data to support informed decision-making in financial markets (Sgammini, 2023). Furthermore, there is a statistically significant linear correlation between the FTSE/JSE index and macroeconomic variables (Banda *et al.*, 2019; Magwedere & Marozva, 2017; Mpofu, 2011). By tracking these indices, stakeholders gain a deeper understanding of market trends, sector performance, and the broader economic landscape. This facilitates more effective investment strategies and regulatory actions.

2.2.2.4 Market sectors

The Global Industry Classification Standard categorised all stocks into ten stock market sectors based on their attributes, such as the industry in which the company operates (Gilbert *et al.*, 2024). The stock market sectors on the JSE were reclassified in March 2000 (Van Rensburg, 2002). The ten sectors include the basic materials, industrials, consumer goods, consumer services, healthcare, telecommunications, utilities, oil and gas, financials and technology sectors (Alovokpinhou & Mbatha, 2022).

Focusing on stocks that exhibit high liquidity and intense responsiveness to news events can significantly improve the performance of a machine learning model designed to predict stock movements (Li *et al.*, 2023; Liu *et al.*, 2023). High liquidity in financial markets facilitates a smoother execution of trades and contributes to more stable price signals (Meldrum & Sokolinskiy, 2025). This is because liquid markets allow for the rapid matching of buy and sell orders, reducing the impact of individual trades on the overall price and enhancing market efficiency (Holden *et al.*, 2014; Rösch, 2021). Additionally, liquidity is critical in minimising transaction costs and promptly ensuring that prices reflect available information (Boudiaf *et al.*, 2024). Furthermore, stocks that are sensitive to news events tend to exhibit timely reactions to new information, which is essential for real-time sentiment-driven predictions (Anastasiu *et al.*, 2019).

The financial sector, comprising banks, insurance companies, and asset managers, consistently exhibits high trading volumes and intense reactions to macroeconomic indicators and regulatory developments (Chellaswamy *et al.*, 2020; Nguyen & Vo, 2024). Therefore, liquidity significantly impacts the performance of South African banks listed on the JSE, which reinforces the sector's relevance for predictive modelling (Mofokeng & Moodley, 2023; Kumalo, 2023). In addition to the financial sector, the consumer goods and industrial sectors on the JSE exhibit signs of high liquidity, as evidenced by consistent investor activity, valuation dynamics, and active trading patterns (Tabot, 2022).

2.3 Stock price determination

The process through which the market price of a stock is established based on various forces of supply and demand, or stock price determination, plays a crucial role in financial markets (Lebepe, 2020). It influences investment decisions, affects market stability and shapes economic trends. Prices fluctuate due to factors such as company performance, macroeconomic conditions, investor sentiment, and external forces (Brijlal & Enow, 2016; Ferreira-Schenk *et al.*, 2024). Understanding stock price movements helps investors assess risk and identify potential opportunities (Gilbertson & Roux, 2006). This section explores the key drivers of stock prices.

Stock prices are directly impacted by the forces of demand and supply (Jha *et al.*, 2024; Mpofu, 2011). At the core of stock price formation lies the interaction between these forces, with buyers and sellers constantly influencing price levels (Masoud, 2013). Prices rise when demand exceeds supply and fall when supply surpasses demand as buyers and sellers adjust to market conditions (Hopman, 2007). A stock exchange traces the supply and demand (Masoud, 2013). Jain (2014) explains that it is one of the most fundamental economic concepts and the foundation of a market economy. Demand reflects the quantity of products that buyers are willing to purchase at a given

price, while supply represents the quantity producers are willing to offer. The relationship between price and quantity-demanded forms the demand relationship, and the link between price and quantity-supplied, defines the supply relationship. The law of supply and demand determines pricing, as high demand allows producers to set higher prices, thereby incentivising production. However, excessively high prices reduce consumer purchases, requiring an optimal balance for sustained sales and profitability. This principle applies to the stock market, where stock price patterns depend on the balance between buyers and sellers.

Investor sentiment, which reflects the mindset of investors, drives demand and supply in the financial market, which influences price changes through optimism or pessimism (Andleeb & Hassan, 2023; Lordkipanidze, 2023). Hanif and Mehmood (2014) explain that investor sentiment is connected to bullish and bearish market trends. Sentiment serves as an excellent indicator of stock price movements as it is directly influenced by the law of supply and demand and shifts between bearish and bullish market trends. Consequently, identifying stocks that are particularly sensitive to sentiment is regarded as an essential component in sentiment-based predictive modelling (Aboutabl *et al.*, 2024). Certain stocks may exhibit stronger price reactions to sentiment fluctuations, thus offering greater potential for sentiment-driven trading strategies (Aggarwal *et al.*, 2023; Anh *et al.*, 2024). The financial sector, characterised by consistently high liquidity and pronounced sensitivity to macroeconomic and regulatory developments, is particularly relevant for predictive modelling on the JSE (Chellaswamy *et al.*, 2020; Kumalo, 2023; Mofokeng & Moodley, 2023; Nguyen & Vo, 2024). Similarly, the consumer goods and industrial sectors exhibit strong liquidity, supported by active trading activity and persistent investor engagement (Tabot, 2022).

2.4 Factors influencing stock prices

Stock prices fluctuate due to a combination of macroeconomic, microeconomic, political, regulatory, and global factors (Almashaqbeh *et al.*, 2021). This is because these elements directly influence market functioning by affecting both the ease of trading (liquidity) and the degree of price uncertainty (volatility), thereby shaping stock price fluctuations (Wang, 2022). Moreover, the connection between macroeconomic conditions, political factors, and share prices provides valuable insights for investors (De Beer *et al.*, 2015). This section covers macroeconomic factors that include GDP growth, inflation, interest rates and exchange rates. It also examines microeconomic and firm-specific factors as well as global events and market shocks.

2.4.1 Macroeconomic factors

Macroeconomic conditions play a fundamental role in shaping market trends and stock price movements (Jha *et al.*, 2024). Inflation, GDP growth, money supply, the unemployment rate, and

interest rates are some of the macroeconomic factors that influence investor sentiment and corporate profitability (Alam *et al.*, 2016; Ergun & Özlen, 2012; Lakra, 2024). Rising inflation erodes purchasing power, increases production costs, and reduces consumer spending, which can negatively impact corporate earnings and stock valuations (Chandrashekar *et al.*, 2018). GDP growth signals economic expansion, thereby fostering business investment and market optimism (Jha *et al.*, 2024). Conversely, economic contraction leads to reduced corporate revenues and lower investor confidence. Interest rates set by central banks affect the borrowing costs and capital flows, which influence corporate expansion and investment decisions (Hsing, 2011; Jhunjhunwala, 2024). Higher interest rates discourage borrowing and increase the cost of capital, while lower rates stimulate investment and economic activity (Caggese & Pérez-Orive, 2022; Mehta *et al.*, 2011). These macroeconomic variables create market cycles that form long-term price trends and volatility (Dsouza *et al.*, 2024; Hsing, 2011).

2.4.2 Microeconomic factors and firm-specific factors

Microeconomic factors refer to the elements that affect the decisions and behaviours of individuals, households, and businesses within an economy. Company-specific factors contribute to stock price fluctuations by reflecting a firm's financial health, operational performance and strategic direction (Gujrati, 2015; Prazak, 2020). Earnings reports, revenue growth, profit margins, company expansion, and balance sheet strength all affect investor expectations and valuation metrics (Hussain *et al.*, 2022). Firms that consistently exceed earnings projections often attract greater investor interest, while financial underperformance can lead to selloffs (Sadeghzadeh, 2018). Competitive positioning within an industry, brand reputation and operational efficiency further shape investor confidence and long-term valuation (Hussain *et al.*, 2022). Microeconomic factors and firm-specific factors are confirmed to influence supply and demand (Gujrati, 2015).

2.4.3 Global events and market shocks

Global events create sudden shifts in stock prices by altering market sentiment and economic expectations (Al-Absy *et al.*, 2024). Financial crises, pandemics, geopolitical conflicts and natural disasters disrupt economic activity, affecting supply chains, corporate revenues and investor behaviour (Attílio, 2024). Market shocks often trigger widespread selloffs as uncertainty increases, leading to heightened volatility (Al-Absy *et al.*, 2024). International capital flows respond to global economic conditions, with investors reallocating assets in response to risk exposure (Attílio, 2024). Commodity price fluctuations, currency exchange rate movements and shifts in global trade dynamics influence corporate earnings and sectoral performance (Kang *et al.*, 2020). While some events create short-term market turbulence, others reshape long-term investment trends by altering global economic structures and financial market dynamics

(Chakraborty *et al.*, 2019). The balance of supply and demand is altered by international events and market shocks (Al-Absy *et al.*, 2024).

Macroeconomic, microeconomic, firm-specific factors, global events, and market shocks all play a significant role in influencing supply and demand dynamics within financial markets (Al-Absy *et al.*, 2024; Gujrati, 2015; Mpofu, 2011). Nevertheless, it is essential to recognise that investor sentiment not only operates alongside these forces but actively shapes the way they are perceived and acted upon by market participants (Wang, 2023). Sentiment influences how macroeconomic data such as interest rates or GDP growth, are interpreted, how firm-specific news like earnings-reports are received, and how global events such as geopolitical tensions are evaluated (Chen *et al.*, 2021; Chen *et al.*, 2021). In this way, sentiment acts as an intermediary lens through which all other factors are filtered, ultimately amplifying or dampening their impact on market behaviour (Andleeb & Hassan, 2023; Hanif & Mehmood, 2014; Lordkipanidze, 2023). Thus, rather than being a separate influence, sentiment is deeply embedded in the market's collective response to a wide range of economic and firm-specific stimuli. Furthermore, investors develop sentiment in response to specific events reported in the news, which might include reports of terrorist attacks, changes in oil prices (Feuerriegel *et al.*, 2015), inflation (Macaulay & Song, 2023), soccer game outcomes (Äijö *et al.*, 2018), businesses, the economy (Eswar *et al.*, 2024), crime, and politics (Yu, 2023).

Given that sentiment plays a critical and integrative role in shaping market behaviour across different levels of influence, the issue of accurately identifying and quantifying investor sentiment becomes essential for understanding and predicting stock price movements. Sentiment analysis offers a systematic approach to capturing this otherwise intangible market force, enabling researchers and practitioners to incorporate behavioural aspects into predictive models. Therefore, it is necessary to explore how sentiment analysis has been applied within the context of stock market prediction, as it provides an important bridge between traditional economic indicators and the behavioural drivers that often govern actual trading decisions.

2.5 Sentiment analysis

Before conducting a comprehensive literature review, it is essential to outline the basic definitions, key concepts, and relevant terminology that form the foundation of sentiment analysis. This section will, therefore, introduce and explain the core ideas and terms related to sentiment analysis, creating a solid foundation for the subsequent literature review.

2.5.1 Introduction

Sentiment analysis is generally categorised into four main approaches: fine-grained analysis, aspect-based analysis, intent-based analysis, and emotion detection (Bakal & Shah, 2022; Chua *et al.*, 2022). These approaches include polarity, emotional detection, and the identification of underlying intentions (Solanki, 2019).

- *Fine-grained sentiment analysis* focuses mainly on identifying specific sentiments based on intensity and emotions in applications, like product reviews and customer feedback analyses (Hoste & Jacobs, 2021). It can be categorised as positive, negative, or neutral (Sudhir & Suresh, 2021). An example can be when the users' sentiments relating to a smartphone are investigated.
- *Aspect-based*, or feature-based sentiment analysis, also focuses primarily on extracting sentiment but applies to specific aspects of a topic (Barbaglia *et al.*, 2022; Chifu & Fournier, 2023). Reverting to the smartphone example, users may be asked to describe their opinions on its battery, camera, screen, and colour.
- In *intent-based sentiment analysis*, a distinction exists between intent and the degree of sentiment within that intent relating to textual data (Komalalakshmi, 2023). For example, the intent of a review could be a query, complaint, command, or desire, each of which can be expressed in levels.
- In *emotion detection*, an analysis extends beyond sentiment categorisation where distinct human emotions are identified and extracted from textual data, and this can be performed at aspect-, document-, or sentence-level (Amirthapriya & Baboo, 2022; Quadri & Selvakumar, 2020). At the aspect-level, each feature of a textual sentence is examined separately, while a piece of text is considered in its entirety at the document-level. Sentence-level analysis involves tokenisation of the text into individual sentences and analysing them separately.

Publications on sentiment analysis research started to increase after the year 2004 and showed a noticeable upsurge in recent years, making it a very young research area (Aider & Palomino, 2022; Dehanka *et al.*, 2022; Khairnar & Kinikar, 2013). The vast majority of research on such analyses is performed on online reviews and social media data retrieved from online social networks in cases where user-opinion information is accessible (Chua *et al.*, 2022; Hernandez-Suarez *et al.*, 2018).

Public opinions shared on social media by users from diverse regions make these platforms valuable sources for sentiment analysis (Abbas *et al.*, 2022). Social media platforms contain vast amounts of raw data uploaded by users, which can include text, videos, photos, and audio clips (Drus & Khalid, 2019). Social networks, on the other hand, are found in the form of websites, computer programs or applications that enable users to communicate and interact by sharing

information, messages, images, and videos (Abbas *et al.*, 2022). They are internet-based social media platforms, also known as the web (Musial & Kazienko, 2013) which is:

“...an open environment for publishing and exchanging data that is constantly developing.” (Aguilera *et al.*, 2019: 116)

Not only is the online environment opinion-rich, but the fast-changing pace and increasing volumes of online information make it popular among academia and professionals in terms of data collection (Sacramento, 2021).

Several emotion theory models are used to identify basic emotions (Avetisyan *et al.*, 2016; Fathalla, 2020), which are also referred to as primary, fundamental or first-order emotions (Ortony, 2022). However, experts are unable to agree on the number, types, or reasons why emotions should be classified as basic (Ortony, 2021). A combination of primary emotions may result in what is often referred to as complex or secondary emotions. For example, a combination of joy and trust may be considered as love (Bourgeois *et al.*, 2019). Numerous case studies have demonstrated that negative emotions lead to poor decision-making, and individuals who experienced more intense feelings achieved higher decision-making performance (Carlson & Cousineau, 2020; Kamalanabhan & Navaneedhan, 2020). Emotions form an elemental part of our cognitive and mental functions and are, therefore, a massive contributor to communication and decision-making (Shakhovska *et al.*, 2020). They ultimately translate to emotional intelligence, which significantly influences thought processes, decision-making, and individual success (Raghvendra, 2024).

Ahsan *et al.* (2016) studied the actions of 225 stock investors and found that their investor decisions were impacted by emotional intelligence. Emotional intelligence is the ability to understand both one's own emotions as well as those of others (Ali *et al.*, 2023) and comprises five dimensions: self-awareness, self-management, motivation, empathy and relationship management. The idea that emotions form the key link between past outcomes and present decision-making is argued in many fields of economic research (Cordes *et al.*, 2023). Therefore, emotions and emotional intelligence play a crucial role in economic decision-making (Aftab *et al.*, 2023; Cabrera *et al.*, 2018; Kamalanabhan & Navaneedhan, 2020). Since emotions and emotional intelligence shape decision-making, incorporating emotion models is vital for realistic behavioural modelling (Fiske & Schacter, 2022). Emotion models span a broad spectrum - from Tomkins's biologically rooted *Affect Theory*, which views emotions as innate drivers of behaviour with evolutionary significance (McIlwain, 2007; Ortony & Turner, 1990), to *Izard's Differential Emotions Theory*, which links emotional development to distinct neural circuits and physiological processes (Fathalla, 2020; Wierzbicka, 2010). The continuum also includes *Frijda's Appraisal Theory*, which conceptualises emotions as evaluations of events relevant to personal goals and values (Avetisyan *et al.*, 2016; Moors & Scherer, 2019).

A tree-structured model called the *Parrott Emotion Classification Model* was developed by Parrott (2001) and founded on Ekman's theory of basic emotions (Ekman, 1992). It categorises emotions into three layers: primary, secondary, and tertiary, according to their intensity and nature (Abbasi & Beltiukov, 2019; Fathalla, 2020). The *Colour Wheel of Emotion Model* was established by Plutchik (2001) and inspired by Darwin's theory of evolution, comprises eight primary emotions considered biologically primitive and is believed to have formed the foundation of numerous models that followed (Abbasi & Beltiukov, 2019; Abu-Salih *et al.*, 2023; Fathalla, 2020; Orts & Vargas-Sierra, 2023). The wheel is organised based on the intensity of each emotion, with the strongest emotions located at the centre and diminishing in intensity as you move toward the edges. The model consists of an equal distribution of positive and negative emotions.

Alaghband and McCarthy (2023) discovered that there is a moderate correlation, exceeding 60%, between emotional sentiment and market movements. Emotional factors significantly influence stock prices, but the emotions of investors also have a cyclical nature, where a period of positive emotion is always followed by a period of negative emotion (Xue, 2023). Cabrera *et al.* (2018) identified a research gap regarding the influence of artificial emotion regulation on the effectiveness of autonomous decision-making in stock trading environments. To address this, a mechanism was proposed that integrates both rational and artificially modelled emotional criteria within artificial investors to enhance autonomous trading decisions. The impact of different emotional sensitivity settings, referred to as emotional bandwidths, and the speed of emotional adaptation, referred to as update rates, were tested to evaluate how emotional regulation affects trading decisions and outcomes. In this context, artificial emotions can be constructed as synthetic representations based on the observation and modelling of human emotional behaviour. This highlights the importance of integrating data related to investor emotion to improve trading strategies and decision-making processes and underscores the lack of research on emotion analysis in stock trading.

Xu *et al.* (2018) and Xue (2023) explained that four moods, which include disgust, joy, sadness, and fear, can be used to predict the price and volume of certain stocks in the stock market. Griffith (2020) also claimed that fear among investors significantly and persistently impacted market returns and volatility. Dardana and Elgeka (2023) argued that happiness and sadness are the strongest emotions. Emotion detection models in stock trading can vary significantly across different datasets, as they often originate from diverse sources, such as social media platforms, which also differ by country (Hajar *et al.*, 2016). Investor sentiment has become a particularly relevant area of research, especially since 2014 (López-Cabarcos *et al.*, 2020).

2.5.2 Sentiment analysis in the stock market

The use of sentiment analysis has become an essential tool for understanding some market dynamics, as investor sentiment influences stock price movements. In financial markets, sentiment analysis can be applied to news articles, social media, and financial reports to gauge market sentiment and predict stock price fluctuations. Various modelling approaches have been developed to incorporate sentiment data into specific trading strategies based on precise criteria. This section explores how sentiment analysis has been integrated into stock market models, highlighting key methodologies and their effectiveness in predicting market trends from 2022.

Significant contributions were made in the field when Akeji *et al.* (2022) proposed a novel stock price prediction approach combining sentiment analysis with deep learning. By integrating the *Dragonfly Algorithm* for hyperparameter optimisation, the model improved the accuracy of stock price forecasts using Twitter/X sentiment data. This algorithm is a method used for optimising expensive and high-dimensional functions by efficiently balancing exploration and exploitation. The results highlighted the effectiveness of the *Dragonfly Algorithm*-based deep-belief network model framework in capturing market sentiment and enhancing predictive performance. Experimental findings showed a prediction accuracy of 94.97%, demonstrating the model's reliability. The approach offered a promising direction for leveraging social media data in financial forecasting.

Chakrabarti *et al.* (2022) explored stock market prediction of future movements in stock prices using sentiment analysis on public opinion during the COVID-19 pandemic, leveraging a Long Short-Term Memory (LSTM) machine learning model and seven different sentiment analysis tools. By integrating sentiment scores from multiple sources, the study achieved high prediction accuracy, with the best results obtained from the linear support vector machine with a stock market forecast accuracy of 98.32%. These findings highlight the potential of combining sentiment analysis with stock market data to enhance predictive models, particularly during periods of high public sentiment like the COVID-19 outbreak.

In a study by Cristescu *et al.* (2022), the impact of sentiment analysis on stock price prediction was investigated, emphasising its role in refining risk management models. Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables, allowing for predictions and insights into how changes in the independent variables affect the dependent variable (Kumari & Yadav, 2018). By integrating sentiment scores into regression frameworks, the study demonstrated that non-linear autoregression models achieved a better fit to the data, with polynomial regressions

outperforming linear ones. The results emphasised the advantages of incorporating market sentiment as a factor in financial forecasting, particularly in uncertain economic conditions.

In a comprehensive literature survey, Gajapathy and Sathyanarayana (2022) analysed how sentiment has been applied in stock market prediction by reviewing 204 documents published between 1994 and 2022. The study identified key journals, highly cited authors, contributing countries, data sources, and prediction techniques. Research articles were sourced from the Scopus database, and a systematic review was conducted using the PRISM framework, with VOSviewer software - the latter was used to generate visual maps which highlighted contributing countries and frequently used keywords. Their findings emphasised the growing importance of sentiment derived from news-feeds, particularly Really Simple Syndication (RSS) feeds, in enhancing stock market predictions. By combining sentiment analysis of RSS feeds with stock market data, they demonstrated that this approach could significantly improve prediction accuracy, reporting a 14.43% improvement compared to standard algorithms, such as decision trees and the moving average indicator. These findings provide necessary support for the inclusion of RSS feed data in stock market sentiment analysis, helping to inform the data collection approach for this study.

The impact of electronic word-of-mouth data on stock trading and market confidence was examined by Lai *et al.* (2022). Their research emphasised the importance of sentiment analysis in multiple languages in forecasting movements in the stock markets. Since most sentiment analyses are conducted in English, texts from non-native English-speaking countries were translated for a multilingual analysis. Unstructured social media data were cleaned and combined with structured trading data and technical indicators. An LSTM model optimised with a genetic algorithm to incorporate multilingual sentiment data, called LSTMGA, was used to predict stock indices and company stock prices. Results showed that the LSTMGA model outperformed other machine learning models, demonstrating its effectiveness for stock market forecasting.

The use of advanced machine learning techniques, particularly recurrent neural networks (RNNs) and LSTM models, was studied by Adhikary *et al.* (2023) to predict stock market trends while incorporating sentiment analysis for more accurate forecasting. By enhancing text classification with the Natural Language Processing (NLP) model BERT (Bidirectional Encoder Representations from Transformers) and its variants, the research demonstrated that combining these technologies improved prediction performance regarding accuracy, precision, and recall (Alsaifi *et al.*, 2025). Recall, also referred to as sensitivity, quantifies the proportion of actual positive instances that are correctly identified by the model, indicating how effectively the model predicts positive cases (Sathyanarayanan & Tantri, 2024). The findings underscored the potential

of machine learning and sentiment analysis in providing more reliable insights for investors, ultimately leading to better-informed decisions in the stock market.

Dewan *et al.* (2023) analysed the interpretability in Artificial Intelligence (AI)-driven sentiment analysis to enhance fairness, accountability, and transparency. The study employed interpretability techniques in NLP-based sentiment analysis by using textual data collected from social media platforms. The lexicon and rule-based sentiment analysis tool, Valence Aware Dictionary for Sentiment Reasoning (VADER), was used to generate heatmaps to provide visual justifications for forecasts. At the same time, locally interpretable model-agnostic explanations offered more profound insights into predictions. Experimental results demonstrated that the proposed system improved interpretability compared to existing approaches at the time.

Although stock prices are influenced by financial news, relationships between words are often overlooked in sentiment analysis. John and Latha (2023) proposed a hybrid RNN combining bidirectional-LSTM, gated recurrent unit, and stacked LSTM to predict stock price trends using stock features and financial news, and the relationship between them. Results showed that integrating sentiment and semantics significantly improved forecasting, outperforming traditional models. The strong potential of deep learning for stock price prediction was highlighted.

Karanikola *et al.* (2023) investigated how integrating sentiment analysis into the feature space of financial time series forecasting could enhance model performance. Through an extensive experimental process, 67 different feature-and-sentiment setups were tested across a range of datasets and evaluated using multiple metrics. A feature-and-sentiment setup is the combination of stock closing prices with sentiment scores. A total of 30 state-of-the-art deep learning algorithms were compared. The study identified a best-performing method based on average rankings and statistical tests, demonstrating that incorporating sentiment-related features can conditionally improve model efficiency, particularly for specific forecast time frames.

Neutrosophic logic introduces the idea of three independent components, namely truth, indeterminacy, and falsity (Smarandache, 2024). This is used in decision-making where uncertainty and conflicting information exist. Abdelfattah *et al.* (2024) presented an advanced stock market prediction framework integrating neutrosophic logic with sentiment analysis and deep learning. Their work demonstrated the effectiveness of handling uncertainty in sentiment classification, leading to more reliable stock movement forecasts. The approach successfully enhanced prediction accuracy by combining sentiment data and historical market information. Refining sentiment analysis through advanced techniques can, therefore, significantly impact the reliability of stock market forecasts.

In a study aimed at assessing the impact of combining sentiment analysis with historical data, Aboutabl *et al.* (2024) developed an AI model for stock price prediction using LSTM and Extreme Gradient Boosting (XGBoost) techniques. XGBoost is a powerful machine learning algorithm that builds an ensemble of decision trees using gradient boosting to improve predictive accuracy, particularly for structured/tabular data (Abduljabbar *et al.*, 2023). They applied their model to the stock data related to Apple, Google, and Tesla. Sentiment scores were computed using NLP and then integrated with historical data. Performance evaluation included the use of Root Mean Square Error (RMSE) to measure the average magnitude of prediction errors, Coefficient of Determination (R^2) to assess the proportion of variance explained by the model and Mean Absolute Error (MAE) to quantify the average absolute error between predicted and actual values. The incorporation of sentiment data together with historical data significantly improved forecasting accuracy when compared to using historical data alone.

Another approach to stock market forecasting involves combining time-series analysis with textual data mining. This was explored by the TextBlob library to link public sentiment with stock price fluctuations. The results demonstrated that LSTM outperformed the one-dimensional convolutional neural network (CNN), particularly for stocks with dramatic price movements, confirming the LSTM's suitability for time-series problems involving sentiment-driven market dynamics. Moreover, the study found that TextBlob provided superior sentiment analysis performance compared to other libraries such as VADER. These findings support the integration of sentiment analysis into stock market forecasting and highlight the importance of selecting appropriate NLP tools to enhance predictive accuracy.

In a study by Bhavsar and Gavhane (2024), the use of data in financial markets was explored, examining its potential for market prediction and profit generation. Sentiment analysis was applied to extract insights from textual content, and NLP was used to analyse a large Twitter/X dataset to assess public opinion on specific financial markets. They aimed to improve the understanding of market trends and investor behaviour. Similarly, Eswar *et al.* (2024) used sentiment analysis together with NLP in the financial sector to assess public sentiment from social media posts¹ about businesses and the economy. The study examined their impact on stock market predictions by classifying sentiments as positive, negative, or neutral. Data for analysing stock price dynamics were collected on platforms such as Twitter/X and Reddit. Prior research was reviewed to assess the effectiveness of sentiment analysis, while addressing key issues such as data accuracy and

¹ Posts refer to user-generated content shared on social media platforms, such as text, images, videos, or links.

market response speed. The study also highlighted the benefits of integrating sentiment analysis with traditional financial metrics to improve stock market forecasting.

The literature highlights the significant role of sentiment analysis in improving stock market prediction models and traces the evolution of its related research and changing techniques in recent years. Advanced approaches, including deep learning models such as LSTMs and RNNs, have improved sentiment classification and predictive accuracy. Studies have shown that integrating diverse sentiment sources such as social media data, news headlines, and financial forums enhances forecasting performance, sometimes outperforming individual sentiment sources. Additionally, methods supplemented by optimisation techniques have demonstrated further improvements in predictive reliability. While sentiment analysis has proven to be effective in stock market forecasting, some challenges remain, particularly regarding uncertainty in sentiment classification, domain-specific language complexities and the ethical implications of AI-driven financial predictions. Nonetheless, the research suggests that refining and integrating sentiment analysis methodologies with traditional financial indicators can lead to more accurate and robust stock market prediction models. Sentiment analysis has the potential to be a critical tool for investors and market analysts, as well as for advancements in AI-driven financial forecasting.

Given the unique characteristics and dynamics of the South African economy, it is essential to examine studies that have specifically applied sentiment analysis to South African stock data. This will provide a more tailored understanding of how market sentiment influences stock prices and ensure that the study's methodology and findings are relevant to the South African financial landscape.

2.5.3 Sentiment analysis in the South African context

This section aims to examine the existing body of work related to sentiment analysis in the South African stock market context. By reviewing relevant studies and methodologies, gaps and opportunities for further investigation can be identified.

South Africa's economic landscape differs significantly from that of developed nations, as a large portion of the population remains in poverty, classifying it as a third-world country (Ukwandu, 2019). The economic impact of the COVID-19 pandemic further exacerbated these challenges, with the economy projected to contract by 3.2% in 2020 (Chaminuka *et al.*, 2022). As a result, strategies and applications that are effective in countries like the United States of America may not directly apply to South Africa due to variations in consumer behaviour, education levels, regulatory frameworks, political structures, and tax policies (Eze, 2024). The use of sentiment analysis in various fields in South Africa reflects the country's distinct socio-economic

environment. Assessing such existing research may provide insight into the methodologies and applications used and offer some insight into how sentiment analysis can be refined to address the specific challenges of stock trading in the local market.

Sentiment analysis has been conducted in South Africa across various domains, from financial and political discussions to corporate communication and social issues. Early research explored sentiment trends in politically charged conversations. Kotzé and Seneka (2018) used a lexicon-based approach to examine Twitter/X discussions about a town called Orania, a controversial Afrikaner community. The study found that discussions were largely negative, often linking the town to broader Afrikaner concerns like farm attacks and land expropriation. Their findings suggested the need for more advanced methods to detect hate speech and abusive language.

“Fees Must Fall” was a South African student-led protest that emerged in 2015, calling for the cancellation of tuition fee hikes and demanding a more accessible, equitable, and affordable higher education system. Khan (2019) expanded the political scope of sentiment analysis by conducting a longitudinal sentiment analysis of social media reports related to this movement. Using lexicon models on over 576,000 *tweets* (short messages shared on the social media platform Twitter/X), the study revealed fluctuations in sentiment corresponding to significant events with notable increases in negative sentiment during critical moments. Additionally, the presence of software robots influencing online discussions stressed the need for further research into robot-driven sentiment manipulation.

The focus shifted to banking sentiment when Clark *et al.* (2020) started analysing online discussions about South Africa’s top five retail banks. Using data from various social media platforms, the study employed logistic regression to track shifts in sentiment and discussion topics to offer firms a methodology to monitor customer perceptions in real-time.

Beyond political discourse, the financial sector became a focal point for sentiment analysis research. Marivate and Terblanche (2021) addressed forecasting gaps in the South African financial sector by applying lexicon-based models and gradient boosting to predict the sentiment of financial news articles. Their study updated an existing sentiment dictionary to better capture financial sentiment in the local context. It demonstrated that traditional lexicon-based and machine learning approaches achieved high predictive accuracies ranging between 84% and 94%. The results revealed a strong correlation between predicted sentiment and share price movements, highlighting the potential of sentiment as an indicator of financial performance. Notably, the study emphasised the limitations of pre-trained sentiment analysers for financial text and recommended future research to expand datasets, explore the lagged effects of stock price

reactions, and contribute towards developing an open-source financial sentiment analysis tool tailored for South African markets.

The COVID-19 pandemic introduced new potential applications for sentiment analysis. Using Twitter/X data, Ahmadi *et al.* (2022) using multiple machine learning models, examined public sentiment towards COVID-19 vaccines in South Africa. Their findings showed that a neutral sentiment dominated, reflecting vaccine hesitancy and that city-specific variations influenced vaccination rates and expired vaccine stock. The findings suggested that Twitter/X posts tagged with geographic information could enhance public health policy by providing community-level insights.

Sentiment analysis was also used to examine multilingual and low-resourced languages in South Africa. Mabokela and Schlippe (2022) assessed the effectiveness of sentiment analysis in identifying social issues in multilingual communities by analysing South African tweets in English, Setswana, and Sepedi. Their findings supported the potential of utilising sentiment analysis on multilingual data to detect societal challenges. Future research aims to expand sentiment classification into additional languages and incorporate engagement metrics like 'likes' (a form of user engagement where individuals show approval or support for a post) and 'retweets' (reposting someone else's tweet).

Brahmbhatt *et al.* (2023) extended sentiment analysis into corporate communication by evaluating the Chairman's Statements, a formal narrative section of annual reports where company leadership discusses strategic direction, performance, and outlook. Using FinBERT, a financial sentiment analysis model based on BERT and trained explicitly on financial texts, the study analysed statements from 700 JSE-listed companies to investigate changes in sentiment before and during the COVID-19 pandemic. The results showed no consistent pattern of sentiment shifts across firms, suggesting that communication strategies during the pandemic varied according to corporate culture and leadership style. The findings highlighted the growing importance of accurately analysing unaudited narrative disclosures, such as the Chairman's Statements, especially during periods of crisis. Future research was recommended to refine sentiment measurement methods and to explore industry-specific trends in corporate communication under exogenous shocks.

Further highlighting language-related sentiment analysis, Masethe *et al.* (2023) examined the sentiment of the application process for government employees seeking membership in the Government Employees Medical Scheme. Their analysis focused on this process using *Sesotho sa Leboa*, a morphologically low-resourced language. Lexicon-based models revealed a predominantly negative sentiment, which emphasised the need for improvements in government

services. Notably, DistilBERT (a lighter version of BERT) achieved 99% accuracy in sentiment computation, reinforcing the effectiveness of advanced NLP techniques in low-resource settings.

Fake news and its emotional impact became another emerging area of research. Gagiano and Marivate (2023) investigated whether sentiment analysis could identify bias in South African news. Using the VADER lexicon on data from the South African Disinformation website, the study found that fake news exhibited exaggerated emotional tones with a particularly strong association with anger. Expanding sentiment analysis to global datasets and incorporating deep learning models could enhance the detection of misinformation.

Environmental, Social, and Governance (ESG) sentiment has also been a recent focus. Nyakurukwa and Seetharam (2023) analysed ESG news sentiment and its influence on stock returns using Bloomberg data. Bloomberg is a global financial technology, data, and media company. Their findings indicated that positive ESG news had a significant impact on the flow of stock information, whereas negative sentiment had a minimal effect. Given the study's focus on mining companies, further research was suggested to explore the influence of ESG news sources on sentiment and stock market reactions.

Together, these studies illustrate the growing impact of sentiment analysis across South Africa's political, financial, social, and corporate landscapes. It is important to note that while sentiment analysis has been applied to various domains in South Africa, its relevance to stock trading strategies has not been extensively explored. Existing research covers stock exchanges such as the Dhaka Stock Exchange (Adhikary *et al.*, 2023), Ghana Stock Exchange (Adekoya *et al.*, 2020), National Association of Securities Dealers Automated Quotations (NASDAQ) (Cristescu *et al.*, 2022), Amman Stock Exchange (Bhardwaj *et al.*, 2015), Indian Stock Market (Bharathi & Geetha, 2017), London Stock Exchange (Kiss & Nemes, 2021), and the New York Stock Exchange (Nguyen *et al.*, 2015). While sentiment analysis in stock market prediction has been widely explored in global markets, a notable gap remains in research focused on the South African market. Moreover, expanding sentiment analysis beyond basic goods to other stock sectors, incorporating deep learning models, and refining sentiment classification through targeted keyword development for each sentiment and emotion, while broadening sentiment classes to include emotion analysis, could significantly enhance trading strategies. Addressing these gaps will improve the understanding of market sentiment and its role in stock price movements, ultimately leading to more effective trading models in South Africa.

The literature on sentiment analysis shows its growing importance in extracting subjective information from text, enabling more profound insights into public opinion, market trends, and user behaviour. Various approaches that include lexicon-based methods and machine learning

techniques have been developed to enhance the accuracy and efficiency of sentiment classification. However, many of the most effective sentiment analysis systems rely on machine learning to handle complex language patterns and adapt to different contexts. Given this relationship, the next section will focus on recent research surrounding machine learning.

2.6 Machine learning

Machine learning algorithms play a critical role in sentiment analysis by identifying patterns in textual data and classifying the underlying sentiment with varying levels of accuracy. The selection of an appropriate algorithm depends on factors such as effectiveness and popularity, among others. This section starts with some fundamentals of machine learning, followed by an exploration of the efficacy and popularity of various machine learning algorithms in the context of sentiment analysis.

2.6.1 Introduction

Machine learning is a subset of AI that aims to develop algorithms and models that allow computers to learn from data and to make predictions or decisions without explicit programming (Goutier *et al.*, 2022). However, before machine learning can be applied to sentiment analysis tasks, it is necessary to first extract meaningful features from raw textual data (Ahmed *et al.*, 2024). Feature extraction techniques aim to transform unstructured text into a structured numerical format that can be understood and processed by machine learning algorithms (Kadir *et al.*, 2021).

Machine learning approaches for sentiment analysis can generally be categorised into shallow learning and deep learning techniques (Deepa *et al.*, 2021). Shallow learning models, such as logistic regression, support vector machines, Naïve Bayes, k-nearest neighbours, and random forests, rely heavily on manually engineered features such as word counts, Term Frequency - Inverse Document Frequency (TF-IDF) scores or predefined sentiment dictionaries (Liao, 2024). In contrast, deep learning models, including CNNs and LSTMs, automatically learn complex feature representations directly from raw text without explicit manual intervention (Hussain, 2024). These deep learning approaches are built upon neural networks, computational architectures inspired by the structure of the human brain, comprising interconnected layers of artificial neurons that progressively extract and transform features from the input data (DeNero *et al.*, 2020).

The four main paradigms of machine learning are supervised learning, where models learn from labelled data; unsupervised learning, which identifies patterns in unlabelled data; semi-supervised learning, which combines both labelled and unlabelled data; and reinforcement learning, where agents learn optimal actions through trial and error in an interactive environment (Dehmer &

Emmert-Streib, 2022). The foundations and principles of machine learning will be discussed in more detail in Chapter 3.

2.6.2 Recent research in machine learning

Machine learning-based emotion detection and sentiment analysis have become an increasingly popular research interest, with various methodologies and evaluation metrics being explored. Alslaity and Orji (2022) systematically reviewed 123 papers and identified supervised learning approaches such as support vector machines and Naïve Bayes as the most widely used, with accuracy being the dominant evaluation metric. The study emphasised the importance of emotion detection in human-computer interaction, enabling adaptive, user-centred systems. Similarly, Harnish *et al.* (2022) explored sentiment analysis as an NLP task, observing the increasing adoption of deep learning and transfer learning techniques. Their research centred on applying sentiment analysis to standard datasets and noted widespread use across industries such as hospitality, government, and education. However, both studies primarily concentrated on generic domains and controlled datasets, whereas in the present research, sentiment analysis will be applied dynamically to financial markets using real-time news sentiment and stock data, with a focus on the South African context.

Sentiment classification was performed by Humpherys and Lazrig (2022) using a dataset of 333 student evaluations in the educational domain. While Naïve Bayes achieved 98% prediction accuracy on predicting sentiment when neutral sentiments were excluded, existing algorithms struggled to classify neutral sentiments accurately. A word-sentiment association strategy improved generalisability by reaching 87% accuracy without requiring pretraining. They identified a need for additional educational sentiment datasets to enhance algorithmic performance.

A broader review by Kandiri *et al.* (2022) examined 24 studies on machine learning techniques for sentiment analysis which revealed that Naïve Bayes, support vector machine, and k-nearest neighbour are the most popular models, with accuracy being the most commonly used evaluation metric. Expanding sentiment analysis to financial applications, Liu *et al.* (2022) proposed the sentiment and indicator-enhanced LSTM model for stock price prediction. Their approach incorporated both traditional data sources, such as historical stock transactions and technical indicators, and non-traditional sources, including financial news and social media posts. Sentiment scores were generated through a convolutional neural network (CNN)-based non-traditional data analysis. Their results demonstrated that integrating multiple data sources improved prediction accuracy, bringing the predicted closing prices of stocks in the China Shanghai A-share market, closer to the actual values. However, their study focused on a relatively small set of companies within a single market. It used sentiment extracted solely from pre-

processed static data. In contrast, the present study aims to dynamically link sentiment to stock price movements across broader market conditions in the South African context to allow for the real-time development of trading strategies.

Research exploring sentiment analysis across different domains, particularly focusing on shallow learning approaches, remains abundant. Anjani *et al.* (2023) which developed a sentiment identification model for user complaints about OVO and DANA (mobile payment platforms based in Indonesia) services on Twitter/X, comparing Naïve Bayes, decision trees, random forest, and k-nearest neighbours models. The k-nearest neighbours model achieved the highest accuracy at 91%, while the other models ranged between 86% and 87%. Addressing multilingual sentiment analysis, Benbrahim and Mercha (2023) examined approaches for handling opinionated textual data across different languages, identifying deep learning as the most effective method. However, progress in this area remains constrained by the limited availability of annotated multilingual datasets. In a study on sentiment analysis of customer reviews from the Bengali e-commerce platform DARAZ, Das *et al.* (2023) evaluated seven machine learning and deep learning models: support vector machine, decision tree, random forest, Naïve Bayes, LSTM, bidirectional LSTM, and one-dimensional CNN. The support vector machine achieved the highest accuracy for both English (82.56%) and Bangla (86.43%) texts, while bidirectional LSTM emerged as the best-performing deep learning model. Their findings contributed to advancing sentiment analysis for underrepresented languages such as Bangla. Jaiswal and Singh (2023) reviewed sentiment analysis methodologies using machine learning models that include linear regression, logistic regression, decision trees, support vector machines, Naïve Bayes, k-means clustering, and random forest, evaluating model performance through metrics such as accuracy, precision, recall, and F1-score, where F1-score represents the harmonic mean of precision and recall, balancing both false positives and false negatives. Similarly, Nawaz (2023) conducted a comparative analysis of machine learning algorithms for sentiment analysis that includes Naïve Bayes, support vector machines, random forest, and neural networks. It was concluded that the best model choice depends heavily on data characteristics, computational resources, and the accuracy requirements of specific applications. Collectively, these studies reflect ongoing advancements in sentiment analysis while also highlighting the importance of careful model selection, dataset availability, and continuous methodological improvements.

Recent work, related to sentiment analysis and deep learning approaches, has highlighted advancements in accuracy and model effectiveness. Azrir *et al.* (2024) focused on sentiment analysis in e-commerce, particularly addressing the challenge of informal language use in

customer reviews. Using Shopee² reviews, the study employed feature extraction techniques such as N-grams³, Bag of Words, and TF-IDF, alongside machine learning models that include Naïve Bayes, support vector machines, gated recurrent networks, and LSTM. The support vector machine with a 2-gram TF-IDF achieved the highest accuracy at 86%, demonstrating the effectiveness of machine learning in handling sentiment analysis despite informal language variations. Expanding on the role of AI, Ghatora *et al.* (2024) compared traditional machine learning models with GPT-4⁴ for sentiment classification. While traditional models performed well on short texts, GPT-4 demonstrated higher precision, recall, and F1-scores for longer, more nuanced texts. This makes it superior for extracting meaningful business insights. Addressing the rapid growth of user-generated content, Mamani-Coaquira and Villanueva (2024) introduced a framework that categorises sentiment analysis models based on structural complexity by reviewing advancements in machine learning and deep learning. The study provided a comprehensive overview of methodologies, challenges, and future research directions. Purushotham *et al.* (2024) proposed a CNN-LSTM model through the integration of BERT with attention mechanisms and evaluated its performance using accuracy, precision, recall, and F1-scores. An overall improvement in sentiment analysis performance was observed, particularly in classifying negative sentiments. However, some inefficiencies were reported when processing highly complex semantic content. Collectively, these studies demonstrate the evolving landscape of sentiment analysis while emphasising the effectiveness of deep learning, large language models, and hybrid approaches in improving sentiment classification across different contexts.

A comparative analysis of various machine learning algorithms for sentiment analysis reveals that no single algorithm consistently outperforms others across all scenarios. Model efficacy depends on data characteristics, computational constraints and the specific application domain. Popular models such as support vector machine, Naïve Bayes, k-nearest neighbour and deep learning approaches that include CNN-LSTM and LSTM, have shown strong performance in different contexts. However, trade-offs exist with some methods excelling in accuracy but struggling with computational efficiency or complex sentiment structures. Selecting an appropriate model requires a nuanced approach that aligns algorithmic capabilities with task-specific needs to maximise sentiment analysis effectiveness. Therefore, to establish a methodology for combining machine learning and sentiment analysis on social media data to develop a dynamic trading

² Shopee is a popular e-commerce platform that operates primarily in Southeast Asia and Taiwan

³ N-gram is a contiguous sequence of n items.

⁴ GPT-4 is a state-of-the-art language model developed by OpenAI.

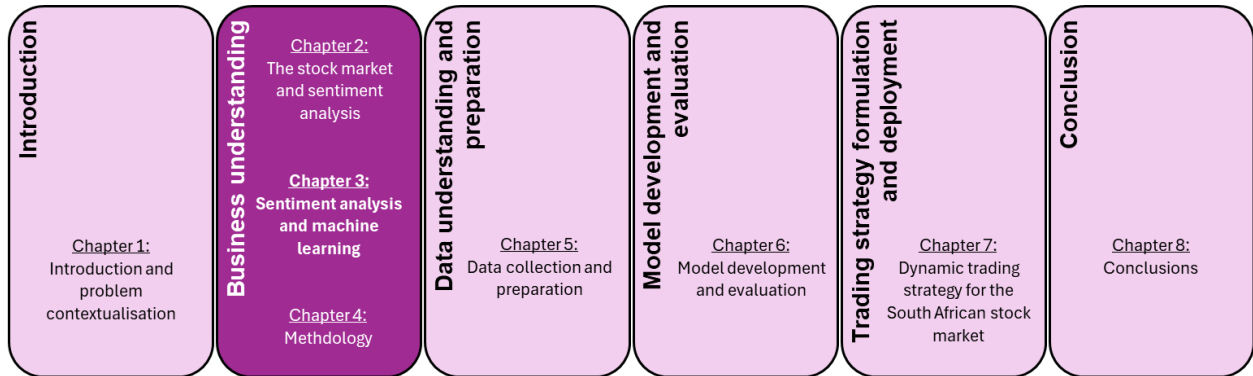
strategy for specific stocks in South Africa, the fundamentals and principles of both these fields need to be investigated.

2.7 Conclusion

This chapter established the foundational understanding necessary for applying sentiment analysis and machine learning to stock market prediction in the South African context. The complex environment in which investor sentiment operates was illustrated by examining the structure of stock exchanges, the unique features of the JSE, and the various factors that influence stock prices. The review identified key gaps in the current literature, specifically in the areas of sentiment analysis and machine learning within the South African market context. Building on these insights, the next chapter will focus on the technical implementation of sentiment analysis and machine learning, detailing the methodologies and models used to predict stock market movements.

CHAPTER 3

Sentiment analysis and machine learning



3.1	Introduction	53
3.2	Fundamentals of sentiment analysis	54
3.2.1	Definition and background	54
3.2.2	Data pre-processing	55
3.2.3	Analysis methodologies	58
3.3	Fundamentals of machine learning	61
3.3.1	Definition	61
3.3.2	Paradigms	61
3.3.3	Approaches	64
3.4	Conclusion	89

3.1 Introduction

The purpose of this chapter is to expand on the principles and processes involved in sentiment analysis and machine learning towards establishing the methodology for this study. Moreover, it forms the second part of phase two in the study structure. Foundational knowledge of these disciplines is necessary to select the correct techniques for the specific application area of social media data and the stock market. Section 3.2 contains discipline-specific terminology and expands on the theoretical processes and different approaches in sentiment analysis. This is followed by a clear definition of machine learning and relevant terminology, elucidating the learning paradigms and in-depth discussions on models commonly used for sentiment analysis applications in Section 3.3. The effectiveness of these models depends on the quality of the data,

the pre-processing, and the selection of suitable algorithms. This chapter, therefore, sets the foundation for establishing the methodology for this research.

3.2 Fundamentals of sentiment analysis

This section defines sentiment analysis and its purpose, emphasising its relevance in fields such as social media analysis. It explores different approaches to sentiment analysis, including rule-based, machine learning-based, and hybrid methods.

Sentiment analysis is the key technique in understanding human emotions and opinions expressed through text. It involves the use of Natural Language Processing (NLP) and occasionally machine learning methods to extract and analyse sentiments from a variety of sources, such as social media, news articles, and financial reports. This section provides an overview of the fundamentals of sentiment analysis, including its definition and background, the pre-processing steps required for accurate analysis, and the various methodologies used to assess sentiment in textual data.

3.2.1 Definition and background

Sentiment analysis is the process of detecting, understanding and interpreting emotions, trends, information, patterns, and the level of satisfaction with the assistance of software (Biswas & Bordoloi, 2023; Chen & Yao, 2021). The general purpose of sentiment analysis is to determine, from language, the polarity of how specific individuals feel, opinions, satisfaction levels, and sentiments toward a particular subject or entity (Aider & Palomino, 2022; Sacramento, 2021). The analysis technique is also known as opinion mining (Aftab *et al.*, 2023; Akhila *et al.*, 2022; Duong & Nguyen-Thi, 2021). While sentiment analysis often draws on user-generated content from social media and online reviews, this study focuses on sentiment extracted from structured news sources to ensure relevance and reliability for stock market prediction (Chua *et al.*, 2022; Hernandez-Suarez *et al.*, 2018). Online users from diverse geographical backgrounds utilise social media platforms to share their views and suggestions, making these platforms highly versatile data sources (Abbas *et al.*, 2022). Consequently, it has become an essential tool for understanding public opinion and emotional responses in various domains (Akhila *et al.*, 2022; Duong & Nguyen-Thi, 2021; Gilbert & Hutto, 2014). These domains include e-commerce feedback, data mining, information retrieval, web data analysis, mining of text, analysis of text, NLP, product analysis, computational linguistics and biometrics (Bhatt & Swarndeeep, 2020; Biswas & Bordoloi, 2023). It is also relevant in fields such as sociology, marketing and advertising, psychology, economics, and political science (Gilbert & Hutto, 2014).

Various challenges accompany the process of modelling and predicting sentiment (Duong & Nguyen-Thi, 2021). These challenges include information overload, redundancy, and relevancy (Chandra *et al.*, 2019). Text data is ubiquitous and, therefore, unstructured and informal because it includes slang terms, emoticons and emojis, abbreviations, hashtags, spelling errors, single characters and improper grammar and punctuation (Halgamuge *et al.*, 2019). These are referred to as “noise” and seen as insignificant and redundant data that may negatively impact a model's performance (Gupta & Gupta, 2019). Therefore, noise removal is one of the essential data-cleaning techniques that must be implemented before accurate analysis can be conducted (Jafarov, 2022).

3.2.2 Data pre-processing

NLP is a tool that enables computers to digest and analyse language as humans do by transforming significant words into numbers (Geetha *et al.*, 2023). The foundation of NLP is pre-processing unstructured and informal text data in preparation for modelling (Parvathi & Ramachandran, 2019). Pre-processing is not solely about improving data quality; it is also about transforming the data into an analysis-ready format that makes it easier to work with (Aloysius & Prakash, 2019; Munk *et al.*, 2017). It reshapes raw data into a proper format by removing irrelevant parts (Akhila *et al.*, 2022). However, different platforms require different pre-processing techniques based on the quality of the data (Aloysius & Prakash, 2019). An example could be, pre-processing newspaper data which is not as extensive as other social media platforms because it usually does not consist of grammatical errors, sarcasm or slang terms (Chandak *et al.*, 2023; Mathew & Mathew, 2023). There are numerous pre-processing techniques, and each of them has an impact on the accuracy and effectiveness of sentiment analysis techniques (Garg & Sharma, 2022). These techniques include removing emojis and emoticons, handling negation, spelling correction, stemming, and platform-specific noise removal on social media, among others. This section explores key pre-processing techniques relevant to this study.

Lemmatisation and stemming are similar in definition in that both shorten a word by transforming it back to its root form. However, lemmatisation adeptly preserves the meaning behind the word by analysing its morphology (Chaniyara *et al.*, 2022). These techniques transform a word (also called a token) into its simplest form by eliminating the affixes that are linguistic elements added to a root or base word to modify its meaning. This process falls under morphological analysis, which broadly refers to the study of the structure of words. However, lemmatisation is a specific type of morphological analysis that preserves the base meaning of words by reducing them to their correct dictionary forms (Ahuja *et al.*, 2019; Keller & Schultz, 2014). The affixes are used to define inflectional words, and by removing them, for example, the word “playing” is transformed into “play” and “studying” is transformed into “study” (Aloysius & Prakash, 2019; De Gruijl *et al.*,

2020). Another example is the set of similar words “great”, “greatly”, “greater”, and “greatest”, which are all converted to “great” to reduce redundancy (Chaudhari, 2020).

Lowercasing is a technique that converts all data to lowercase and is used and recommended by various researchers (Cao *et al.*, 2020; Duong & Nguyen-Thi, 2021; HaCohen-Kerner *et al.*, 2020), for example, the words “tired” and “Tired” are identical in dimensionality, definition, and spelling: computers, however, interpret these words as dissimilar due to the difference in casing. Furthermore, social media users are not particularly attuned to capitalisation and formality; therefore, lowercase text reduces the complexity of the text (Aider & Palomino, 2022). Although it is the simplest pre-processing method, it improves model accuracy considerably (Chaniyara *et al.*, 2022).

Eliminating irrelevant elements such as special characters, stop words, and mentions (when a user directly references another user, brand, or account) is known as *noise removal* (Ahmad *et al.*, 2025). Several improvements are possible:

- Numbers carry no sentiment or emotion and need to be removed to refine and simplify textual data for feature extraction (Jianqiang & Xiaolin, 2017). Number removal is typically performed after the emoticons and the emojis translation, because some of them may contain numbers (Duong & Nguyen-Thi, 2021).
- Punctuation shares parallels with numerical values in the sense that it also lacks sentimental value (Bielikova *et al.*, 2018). Alroobaea *et al.* (2022) state that punctuation can make up between 40% and 50% of textual data. Not only is punctuation insignificant, but a computer is also unable to understand and interpret it (Chaniyara *et al.*, 2022).
- Some raw datasets and previous pre-processing techniques leave extra spaces, which may confuse a model during analysis (Alroobaea *et al.*, 2022). Additionally, it enlarges the field size of the text without adding sentimental value to the data (Abayomi-Alli *et al.*, 2023). For this reason, it is a trivial but significant step in pre-processing to remove excess whitespace in the data (Jiang *et al.*, 2020).
- Stop words are frequently occurring words that have a minimal impact yet discriminative power on the sentiment of a sentence (Adhinata *et al.*, 2023). Parts of speech that fall under the definition of stop words are articles, prepositions and pronouns (Geethangili & Suresh, 2022). Examples are “an”, “and”, “to”, “it”, “the”, “which”, “through”, “in”, “of”, “their” and “this” (Chandak *et al.*, 2023). Garg and Sharma (2022) argued that the removal of these words is compulsory in the successful and precise analysis of sentiment. Moreover, removing stop words improves overall system efficiency by reducing dimensionality and computational cost (Alroobaea *et al.*, 2022). Programs such as Python have built-in dictionaries with stop words,

but they are not predefined and can be improved upon and updated to achieve better accuracy (Garg & Sharma, 2022).

- Special characters or symbols such as “@”, “#”, “*”, and “\$” consume a significant amount of computer memory and prove not to be very important (Alroobaea *et al.*, 2022; Chaniyara *et al.*, 2022; Hargreaves & Ridhwan, 2021). The removal of these symbols is crucial, as they contribute significantly to the noise in the data (Parvathi & Ramachandran, 2019).
- The research on sentiment analysis in languages other than English is limited, as it remains a challenging problem to construct a language-specific sentiment dictionary, which is a highly resource-intensive process (Alfred & Chekima, 2016; Cambria *et al.*, 2016). Model accuracy has been shown to increase when non-English words are transformed or removed from the analysis (Devi *et al.*, 2020; Meher *et al.*, 2020).

Part-of-speech tagging, also known as grammatical tagging, is a classification method that assigns a tag to each word in a sentence (Chandak *et al.*, 2023; Chen & Yao, 2021; Chiche & Yitagesu, 2022). The understanding of the intended meaning of a sentence improves when part-of-speech tags are assigned (Khatter *et al.*, 2023). The classification of tags is segmented into grammatical categories such as adjectives, articles, nouns, and verbs (AlKhatib *et al.*, 2021). The word itself doesn't solely determine its tagging, but it also relies on factors such as its placement, neighbouring words, context, and their respective part-of-speech tags (Li *et al.*, 2021). By performing this step, the potential accuracy of sentiment classification increases (Cheng *et al.*, 2020).

Tokenisation is the technique used to split, for example, paragraphs into sentences or sentences into words (Das *et al.*, 2023; Hidayatullah & Ma'arif, 2017). The formal name for a set of sentences is “corpus” and for the words are “tokens”, hence the name tokenisation (Alroobaea *et al.*, 2022; Choudhry *et al.*, 2023). Breaking a corpus into tokens is the most popular and widely used pre-processing technique in sentiment analysis (Chandak *et al.*, 2023). Ozcelik *et al.* (2023) added that other types of tokenisation include character-level tokenisation and morphological-level (sub-word) tokenisation. The former splits words into their smallest parts; for example, “faster” is tokenised into “f”, “a”, “s”, “t”, “e”, and “r”. The latter splits suffixes and word stems that are semantically more meaningful; for example, “faster” is tokenised into the tokens “fast” and “er”. Tokenisation aids in standardising data by dividing it into smaller units, enabling efficient processing and analysis (Das *et al.*, 2023).

In summary, sentiment analysis can be approached through lexicon-based, machine learning, or hybrid strategies. Each of them offers distinct advantages depending on the complexity and nature of the data. As the choice of approach directly influences model performance, it is essential to apply appropriate data pre-processing techniques and analytical methodologies to ensure

reliable sentiment classification. The following section outlines these preparatory steps and methodological considerations in detail.

3.2.3 Analysis methodologies

Sentiment analysis can be approached in several ways, each offering distinct advantages depending on the context and available data. There are three primary approaches: the lexicon-based (dictionary-based) approach, the machine learning approach, or the hybrid approach (Shinde, 2021). These are visualised in Figure 3.1.

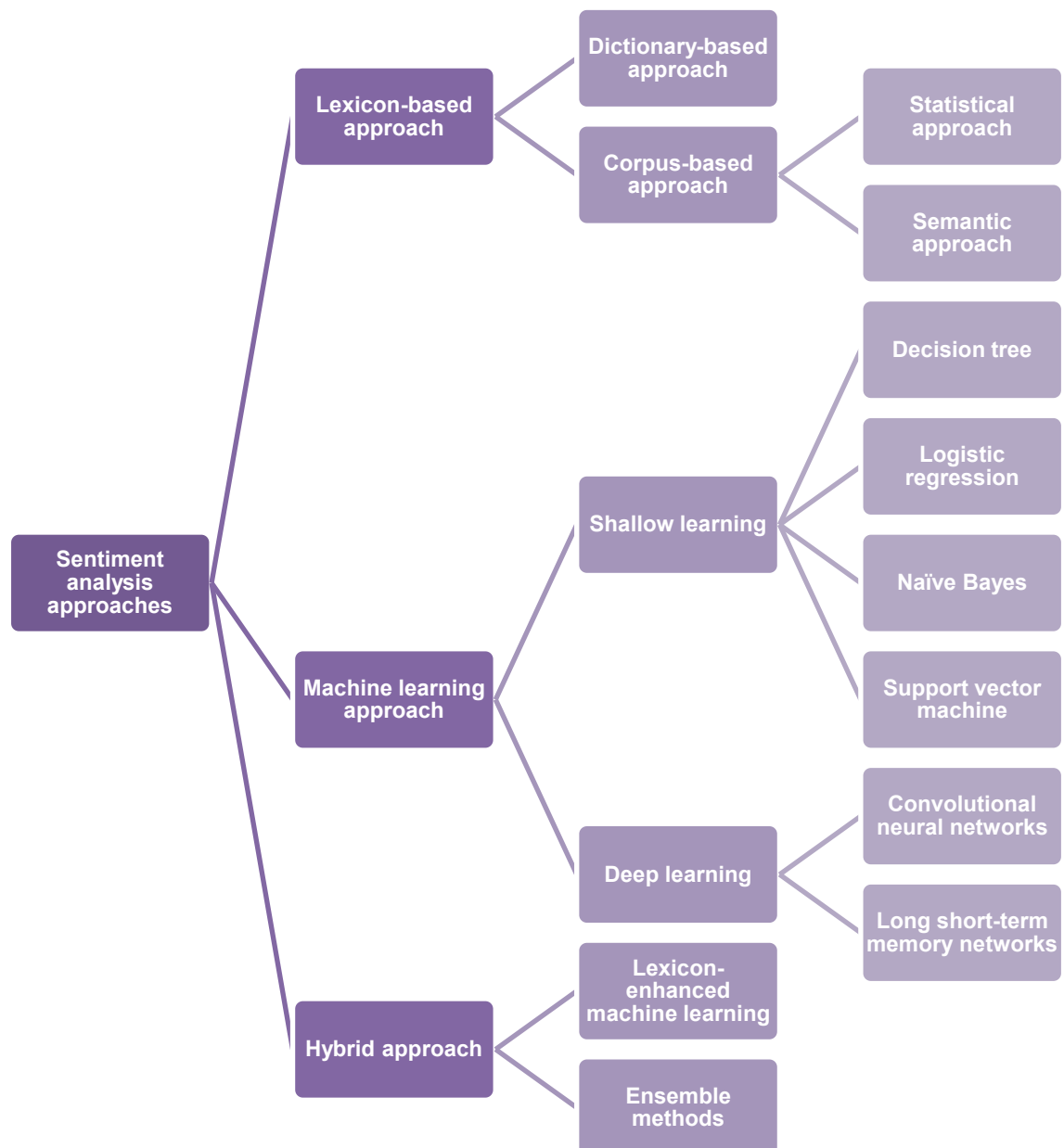


Figure 3.1: Visualisation of sentiment analysis approaches

The *lexicon-based* approach is also known as a rule-based approach because it relies on a set of rules to achieve a predicted sentiment classification (Humpherys & Lazrig, 2022). The rules are used to identify subjectivity, polarity, or opinions (Solanki, 2019). This approach relies on predefined word lists to determine sentiment polarity and can utilise a dictionary with lexicons or a corpus that refines sentiment scores using statistical or semantic techniques. In the case of a corpus, extensive collections of text (corpora) are used to derive sentiment or category associations through statistical or machine-learning methods (Carvalho & Keivandarian, 2023). The range of the lexicon allows it to recognise distinct opinion expressions, yielding better results across different domains. However, its performance depends on many factors, and it cannot be used as a stand-alone analysis due to the challenge of covering every term comprehensively (Liu *et al.*, 2024). Statistically, the recurrence of positive and negative sentiments is considered while semantically, sentiment values can be assigned to words by identifying their synonyms and antonyms (Geethangili & Suresh, 2022). The overall sentiment of a sentence or document is calculated by adding or averaging the values (Nandwani & Verma, 2021). While practical and straightforward, this approach is computationally inefficient, time-intensive, heavily dependent on human input, and performs poorly when dealing with unstructured or informal textual data (Jaffry *et al.*, 2021). When a dictionary is used, predefined lists of words (dictionaries or lexicons) that are associated with specific sentiments or categories are utilised (Butcher *et al.*, 2024). Its advantage is that no prior information about data is needed, as it provides positive results for specific domain vocabulary and quick access to word definitions, but it cannot recognise opinion terms that are not yet included in the dictionary (Liu *et al.*, 2024).

The *machine learning* approach consists of machines that learn how to analyse data to make accurate predictions without any human input or interruption (Bhatt & Swarndeep, 2020). Albaldawi and Alnuaimi (2024) explain that it employs classification techniques to categorise data and regression techniques to predict continuous numerical values. With this method, a dataset is typically split into three parts for training, validation and testing purposes (Burzykowski *et al.*, 2023). This is an autonomous process that is based on machine learning algorithms that train a classifier⁵ from data that comes with labels or has been labelled manually (Geethangili & Suresh, 2022). The training dataset provides the information used to train the model by supplying the characteristics of various instances (Amirkumar *et al.*, 2024). The validation set is used to tune the hyperparameters (Burzykowski *et al.*, 2023). The testing dataset is then used to evaluate the model's performance and determine how well it has been trained (Forghani *et al.*, 2020).

⁵A classifier is a type of **model** that is trained to **assign labels or categories** to data points based on their features.

Machine learning approaches produce highly accurate classifiers compared to rule-based methods; however, the models are often not explicable to humans, making it difficult to diagnose errors and fine-tune a system (Avacharmal, 2024). Different machine learning algorithms yield varying levels of accuracy due to algorithmic complexity and data characteristics (Aftab *et al.*, 2023). The advantage of this type of model is its adaptability to create trained models for specific purposes and contexts (Dahiya *et al.*, 2022). However, it is considered inefficient for new data because obtaining labelled data can be costly or prohibitive (Arablouei *et al.*, 2024). Machine learning comprises shallow learning methods such as decision trees and logistic regression machines, and deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Ankita *et al.*, 2023).

The *hybrid* approach is a combination of the lexicon-based and machine-learning approaches (Hattingh *et al.*, 2023). It is a combination of elements of both of these methods to enhance accuracy and robustness in sentiment classification (Aftab *et al.*, 2023; Geethangili & Suresh, 2022). Researchers agree that the hybrid approach yields better results than the lexicon and machine learning approaches individually (Begum *et al.*, 2024; Bonta *et al.*, 2019; Kapočiūtė-Dzikienė *et al.*, 2013; Musial & Kazienko, 2013). This is because the hybrid approach leverages the strengths of both the lexicon and machine learning approaches while addressing their limitations (Collada-Pérez *et al.*, 2011; Nandwani & Verma, 2021). It combines the precision of lexicon-based sentiment analysis with the contextual understanding of machine learning models, enabling sentiment detection at the concept level with reduced sensitivity to topic changes (Bashar, 2022; D'Andrea *et al.*, 2015). Furthermore, the hybrid approach is more robust in handling variations in language, slang, and evolving sentiment expressions, making it highly suitable for real-world applications where sentiment can be nuanced and multifaceted (Kabir *et al.*, 2024; Vishwakarma & Yadav, 2020).

The choice between rule-based, machine learning and hybrid models depends on the specific requirements of the sentiment analysis task, which includes data availability, desired accuracy, and interpretability. Each approach offers distinct advantages and trade-offs, making it essential to consider these factors when selecting the most appropriate method for a given application. In this study, a combined (hybrid) approach is adopted to leverage the strengths of both lexicon-based and machine learning techniques, aiming to improve accuracy while maintaining interpretability. As a result, it is necessary to explore machine learning methods in greater depth, as discussed in Section 3.3.

3.3 Fundamentals of machine learning

Machine learning plays a vital role in enhancing decision-making in financial markets. This section begins with a definition of machine learning and an examination of various machine learning models, including both conventional shallow learning models and deep learning models, which are essential when adopting a hybrid approach to sentiment analysis.

3.3.1 Definition

As a branch of artificial intelligence (AI), machine learning is essential in sentiment analysis, enabling the extraction of insights and information from opinion-based data (Abbas *et al.*, 2022; Ahmed *et al.*, 2023; Tiwari *et al.*, 2018). The concept was first established as an academic discipline in 1956 at the Massachusetts Institute of Technology (Haenlein & Kaplan, 2019; Pan, 2016). Moreover, it is closely related to and overlaps with computational statistics and data science (Asongo *et al.*, 2021). As a result, machine learning is defined as the scientific discipline that involves the development of algorithms and statistical models that enable computer systems to perform specific tasks without being explicitly programmed (Abbas *et al.*, 2022; Mahesh, 2018; Tiwari *et al.*, 2018).

Understanding machine learning involves identifying two core aspects: the nature of the learning task and the implementation of the learning process. The first relates to the learning paradigm, discussed in the next section, and the second refers to the structure of the learning models, reviewed in the subsequent section.

3.3.2 Paradigms

Machine learning problems are classified into four types, or paradigms, based on how a model learns from data and the type of feedback produced. The four paradigms are supervised, unsupervised, semi-supervised and reinforcement learning (Asongo *et al.*, 2021; Hattingh *et al.*, 2023) and their functions are illustrated in Figure 3.2.

In *supervised learning*, the objective is to learn from labelled data, where each training example includes both the input and the correct output (Asongo *et al.*, 2021). The algorithms guide the training process using example pairs, each consisting of text and its corresponding target (Hattingh *et al.*, 2023). Labelled training data is used to generalise and classify sentiments into new or output data (Nawaz, 2023). Supervised learning is a task-driven approach conducted when specific goals to be achieved are predefined for a given set of inputs (Ngene *et al.*, 2021).

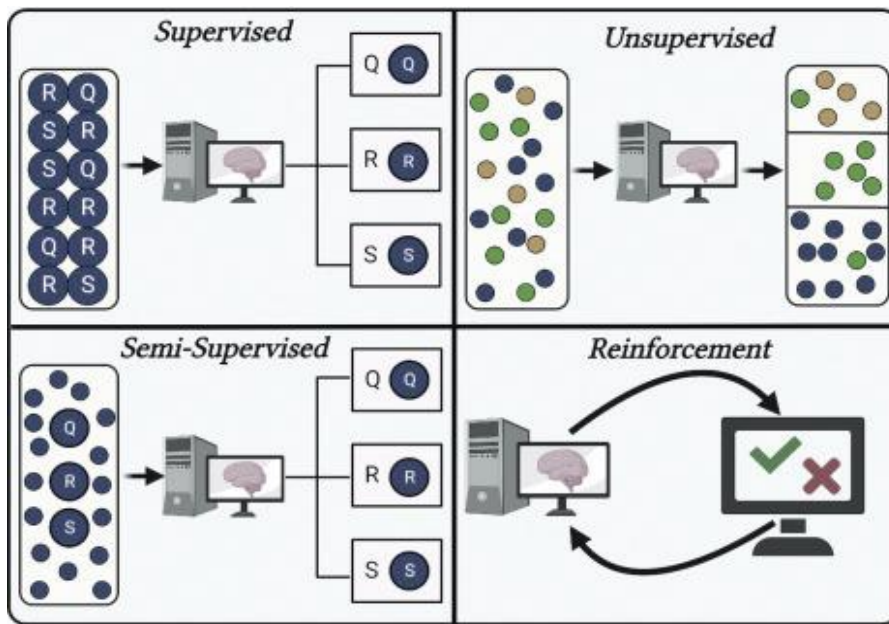


Figure 3.2: The functions of the four common paradigms of machine learning (Camps *et al.*, 2022)

The method offers clear insights into class distributions and is effective for classifying and predicting numerical targets, but it struggles with complex tasks such as understanding contextual relationships in language or processing high-dimensional data, lacks autonomous feature discovery, and requires significant computation time for large datasets (Marini *et al.*, 2021). Supervised algorithms that are especially popular for use in sentiment analysis include decision trees, k-nearest neighbours, logistic regression, Naïve Bayes, random forest, and the support vector machine (Aftab *et al.*, 2023; Asudani *et al.*, 2023; Duong & Nguyen-Thi, 2021; Haddi *et al.*, 2013; Jin *et al.*, 2020; Tiwari *et al.*, 2018).

Unsupervised learning involves training algorithms that find patterns in clusters of data without the use of a pre-defined target or label (Hattingh *et al.*, 2023). This means that training data is neither classified nor labelled, and manual data tagging can be time-consuming (Ahmed *et al.*, 2024; Asongo *et al.*, 2021). Unlabelled datasets are analysed through a data-driven process that operates without the need for human intervention (Ngene *et al.*, 2021). This process can uncover hidden patterns and opportunities that are valuable for businesses, such as fraud detection and identifying new ventures. However, it can be costly, may require human interpretation, and often produces less accurate results with questionable usefulness (Marini *et al.*, 2021). Techniques such as clustering, dimensionality reduction, topic modelling, anomaly detection and association rules are commonly used in this paradigm (Reddy *et al.*, 2018). However, recent developments in machine learning have introduced new perspectives on representation learning, with self-supervised learning emerging as a complementary approach that builds upon principles from

unsupervised learning (Calvo *et al.*, 2025; Dutta & Raj, 2023). To mitigate constraints associated with limited labelled data, self-supervised learning reduces dependence on manual annotation by leveraging the inherent structure and patterns within unlabelled data (Diciotti *et al.*, 2025). This means that rather than identifying clusters or latent structures without explicit targets, self-supervised learning defines proxy prediction tasks directly from the data itself, thereby enabling models to learn meaningful representations (Taherdoost, 2024). These representations are typically general and task-agnostic, and are commonly used for pre-training models before fine-tuning on specific downstream applications (Cole *et al.*, 2025).

The hybridisation of both labelled and unlabelled data during training defines *semi-supervised learning* (Reddy *et al.*, 2018). It is essentially a blend of the previous two methods (Ngene *et al.*, 2021) and extensively applied in machine translation, fraud detection, data labelling, and text classification (Marini *et al.*, 2021). This approach enables the utilisation of vast amounts of unlabelled data alongside typically smaller sets of labelled data in various applications as sentiment analysis, fraud detection, medical diagnosis, and customer behaviour prediction (Hoos & Van Engelen, 2020). Although this method is easy to understand and reduces the need for annotated data, it can lead to unstable results and low prediction accuracy (Marini *et al.*, 2021).

Reinforcement learning is a type of machine learning where an algorithm learns by receiving rewards for correct actions and penalties for incorrect ones (Hattingh *et al.*, 2023). It uses an environment-driven approach, allowing machines and software agents to learn the best actions to take in different situations through trial and error, for example: a trading algorithm that learns to buy or sell stocks based on market conditions and gives feedback from profit or loss outcomes (Asongo *et al.*, 2021; Ngene *et al.*, 2021). The algorithm relies on feedback to identify which actions lead to the highest rewards (Ali *et al.*, 2022). Reinforcement learning, on the other hand, is effective for solving complex, long-term problems and mimics how humans learn from experience; it also requires large amounts of data and computational power. If not managed properly, it can become overwhelmed by too many possible states, which could reduce its effectiveness (Marini *et al.*, 2021).

The output generated by a machine learning model can be either categorical, which is effective for classification purposes, or numerical, which gives values used for analyses like regression (Hasib *et al.*, 2022). Figure 3.3 shows the difference between the two types of output.

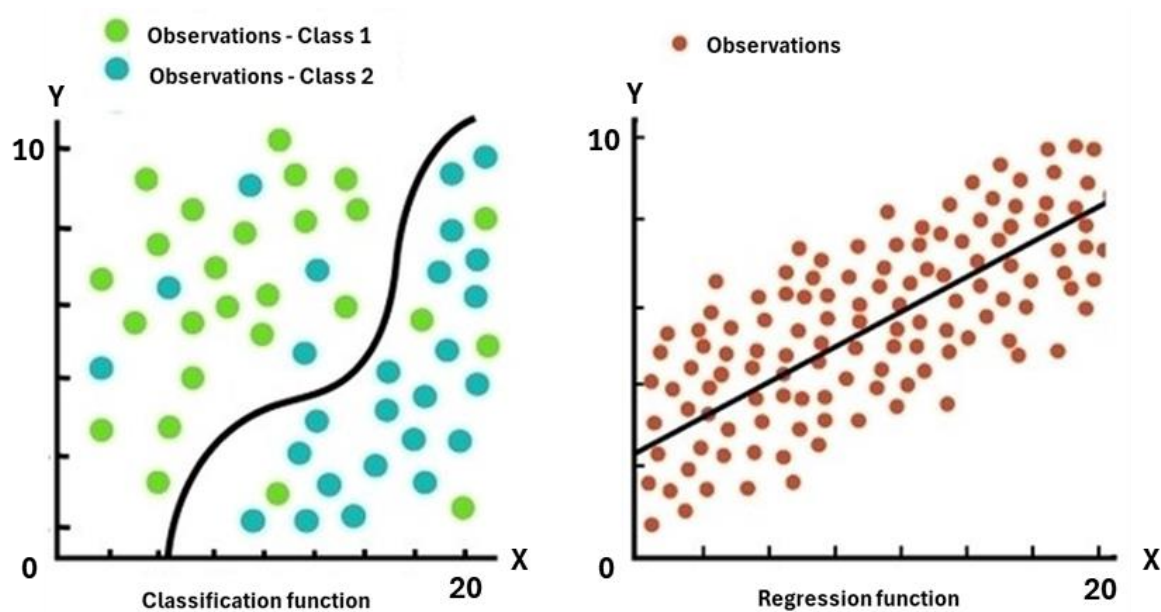


Figure 3.3: Comparison of the output of classification vs regression analyses
(Belkacem, 2021)

During classification, data points can be divided into two categories, typically represented by different colours (e.g., blue and green). The solid black line represents the decision boundary that the model used in these cases to separate the categories. Classification aims to assign new data points to one of these predefined categories based on their features. However, regression is a technique used to understand the relationship between variables, specifically how one or more independent variables affect a dependent variable or output. The diagonal line represents the best-fit line calculated to predict the dependent variable's (Y) value based on the independent variable (X). The key distinction lies in their output: classification outputs discrete labels, while regression outputs continuous values.

While learning paradigms define the type of task and data involved in machine learning, the algorithms are further distinguished by the complexity of their architectures, commonly referred to as shallow and deep learning (Asongo *et al.*, 2021; Haneczok & Piskorski, 2020; Hattingh *et al.*, 2023).

3.3.3 Approaches

The function of a machine learning algorithm is to mathematically transform input data into meaningful outputs by learning from recognisable patterns in the data. Machine learning approaches can be broadly categorised into shallow and deep learning.

3.3.3.1 Shallow learning

This category of learning refers to a technique with a single stage of transformation, or layer, where data is transformed from input features directly to output predictions without involving multiple hidden layers or complex hierarchies. The purpose of such an algorithm is to extract patterns from the input data using weighted connections and mathematical functions. Due to their limited depth, such approaches can effectively learn simple patterns but struggle with complex, hierarchical relationships. While they require fewer computational resources and are easier to train, they often lack the capacity to capture deep representations, making them less effective for tasks involving large and highly unstructured data (Arkhangelskaya & Nikolenko, 2023; Haneczok & Piskorski, 2020; Taherdoost, 2023). Examples of shallow learning approaches include decision trees, the k-nearest neighbour algorithm, logistic regression, a Naïve Bayes classifier, the random forest method, and support vector machines, among others (Abaye *et al.*, 2020).

3.3.3.1.1 Decision tree

The decision tree classifier is a form of supervised learning that works by splitting data into branches based on decision rules derived from input features, constructing a tree-like structure (Kulkarni *et al.*, 2022). At each node in the tree, the algorithm selects the feature that best separates the data using criteria such as Gini impurity or information gain. These splits continue, creating branches until a predefined depth is reached or further splitting no longer adds value.

There are three types of decision tree algorithms, namely Iterative Dichotomiser 3 (ID3), Classification and Regression Trees (CART) and Classification 4.5 (C4.5) (Hasib *et al.*, 2022). ID3 builds trees by selecting attributes that maximise information gain, while CART creates binary trees by splitting the data at each node based on the best feature that minimises impurity (Khairani *et al.*, 2023). The C4.5 algorithm is an extension of the ID3 algorithm as it improves it by handling both continuous and discrete attributes and caters for pruning (Cherfi *et al.*, 2020). Decision trees are suitable for regression (where the target variable is numeric) and classification modelling (with the target variable categorical) (Elbagir & Yang, 2019). The model is simple to build and requires neither a large dataset nor extensive training time (Liu *et al.*, 2024). Extensive data preparation and pre-processing are also not required; however, it is unstable because it is susceptible to changes in the values (Jaiswal & Singh, 2023). A single tree might excel in prediction but may exhibit signs of overfitting (Kanozia *et al.*, 2023). Overfitting occurs when a tree becomes too complex, capturing not only patterns in the data but also some noise or random fluctuations. Figure 3.4 presents the typical structure of a decision tree algorithm.

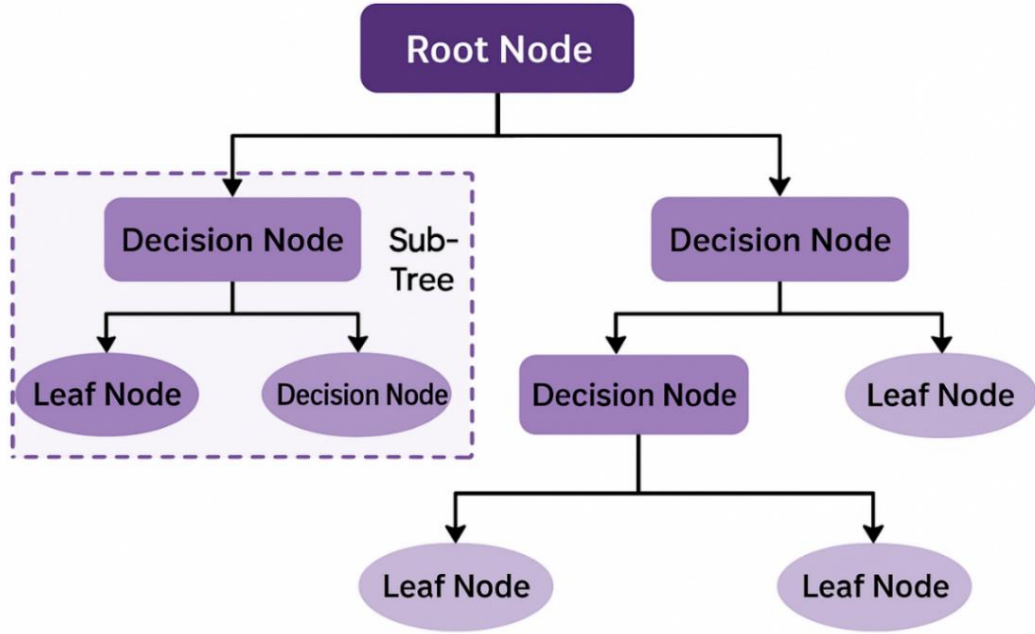


Figure 3.4: Visualisation of the decision tree algorithm (Abdulazeez & Jijo, 2021)

A tree consists of a root, decision nodes, branches and leaf nodes (Abdulazeez & Jijo, 2021). The root node is the topmost node representing the entire dataset; it splits into two or more homogeneous sets. The decision nodes represent where the data splits based on a variable attribute. Branch nodes are intermediate between the root and leaf nodes, representing decisions to further split the data. These nodes continue the process of decision-making until reaching the leaf nodes. Leaf nodes are terminal nodes that predict the outcome or the final decision. The three most common node-splitting criteria are the Gini impurity index, information gain and the Chi-squared measure. Each criterion evaluates the quality of a split based on how well it separates the data.

The *Gini index* is a measure of node impurity or diversity. It can be computed as

$$Gini(p) = 1 - \sum_{k=1}^K p_k^2, \quad (3.1)$$

where K represents the number of classes and p is the vector of class proportions at a given node, with p_k denoting the proportion of records in class k that satisfy the current split condition (Cao *et al.*, 2021). Within a decision tree, each potential split at a node is evaluated using a criterion such as the Gini index, which serves as a proxy for a loss function. A loss function quantifies the difference between the predicted value and the actual value for a single data point.

Information gain assesses the reduction in entropy when splitting a dataset based on a particular attribute. Entropy, which measures disorder or impurity, is defined by

$$Entropy(L) = - \sum_{i=1}^j p_i \log_2(p_i), \quad (3.2)$$

where L typically represents the set of data instances (or a node's dataset) at a particular point in the decision tree and p_i is the proportion of instances in class i out of j total classes (Tangirala,

2020). The information gain ratio then determines the effectiveness of a feature f in reducing uncertainty in the form

$$IG(L, f) = Entropy(L) - \sum_{v=1}^V \frac{|L^v|}{|L|} \times Entropy(L^v), \quad (3.3)$$

where $|L^v|$ represents the size of the subset formed by splitting on the feature f , while V denotes the number of distinct values the feature can take (Tangirala, 2020). A higher information gain suggests a more valuable feature for classification.

Beyond impurity-based measures, the *Chi-squared* criterion evaluates the statistical significance of a split. It compares observed and expected frequencies within a dataset according to

$$X^2 = \sum_{i=1}^N \frac{(O_i - E_i)^2}{E_i}, \quad (3.4)$$

where N is the number of label classes and O_i and E_i denote the observed and expected frequencies for each category i , respectively (Patel & Upadhyay, 2012). A higher Chi-squared value indicates a greater discrepancy between the observed and expected frequencies, making it useful in hypothesis testing and feature selection.

The decision tree classification technique involves two phases, namely tree building which is conducted in a top-down manner and tree pruning, which is performed from the bottom up (Jebaseeli & Kasthuri, 2020). The tree-building process involves the creation of nodes and branches that are pruned either during or after construction to prevent overfitting (Dai *et al.*, 2016). Pruning in machine learning involves removing unnecessary branches or nodes that contribute little to the overall accuracy (Ghodasara, 2021). Another method that prevents overfitting is hyperparameter optimisation (Elgeldawi *et al.*, 2021). This is the process of optimising external configurations of a machine learning model to enhance its performance (Cerri *et al.*, 2024) and are discussed in detail in Chapter 6.

Decision trees construct explicit models by learning a series of if-then rules, whereas other algorithms employ different approaches. The k-nearest neighbour method forgoes training but requires more time to learn.

3.3.3.1.2 The k-nearest neighbour algorithm

The k-nearest neighbour algorithm classifies new instances of data according to the majority label (like the average) of the k most similar instances in the training data (Ahamed *et al.*, 2022). Therefore, the “ k ” refers to the number of nearest data points, or neighbours, considered when making a prediction (Gide *et al.*, 2022), enabling the algorithm to swiftly react to any input changes that happen during real-time use (Defriani *et al.*, 2023). The basic function of the k-nearest neighbour approach is depicted in Figure 3.5.

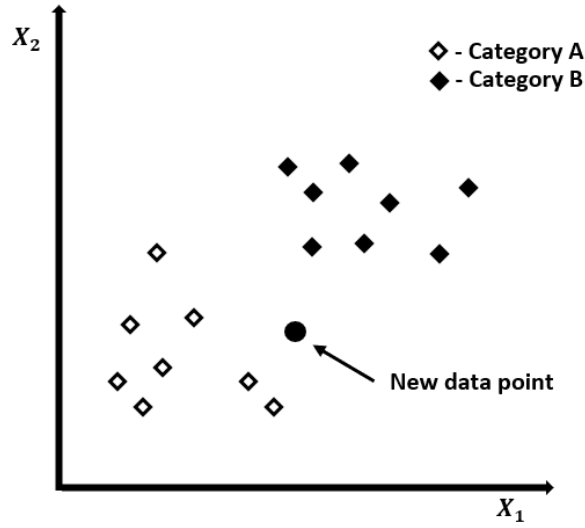


Figure 3.5: Visualisation of the k-nearest neighbour algorithm (Bansal *et al.*, 2022)

Figure 3.5 illustrates how the k-nearest neighbour algorithm uses the distance between new data points and categories to make predictions. Data points are plotted based on two features, X_1 and X_2 . The points are grouped into Category A and Category B. A new data point is introduced, and the k-nearest neighbour algorithm determines its category by analysing the nearest neighbours (categories) and the distance between the new point and its nearest neighbours. The new data point is closest to a group of points from Category A, suggesting that it is likely to be classified in the same category. This model employs the direct distance metric to determine the appropriate class. Other popular metrics that can be used to calculate the distance between two data points include Euclidean, Manhattan, Minkowski, Chebyshev, Cosine, Jaccard, and the Hamming distance metric (Bhat *et al.*, 2022).

For two data points i and j , in an m -dimensional space, with x_{in} and x_{jn} corresponding to the values of the n^{th} dimension, the distance d_{ij} between the two points is calculated as follows:

- The *Euclidean* distance,

$$d_{ij} = \sqrt{\sum_{n=1}^m |x_{in} - x_{jn}|^2}, \quad (3.5)$$

provides an intuitive measure of spatial separation (Abdullah *et al.*, 2021; Pamuji, 2021).

- Another widely used metric is the *Manhattan* distance,

$$d_{ij} = \sum_{n=1}^m |x_{in} - x_{jn}|, \quad (3.6)$$

which measures how far apart the two data objects are in the multi-dimensional space, considering only horizontal and vertical distances, like navigating a grid-like path in a city, hence the name (Bhat *et al.*, 2022).

- The *Minkowski* distance is

$$d_{ij} = \sqrt[p]{\sum_{n=1}^m |x_{in} - x_{jn}|^p}, \quad (3.7)$$

with $p = 1$ for the Manhattan distance and $p = 2$ for the Euclidean distance. It generalises both Euclidean and Manhattan distances through the parameter p , which controls the order of the metric (Alfeilat *et al.*, 2019).

- Also referred to as chessboard distance, the *Chebyshev* distance,

$$d_{ij} = \max_i |x_{in} - x_{jn}|, \quad (3.8)$$

identifies the most significant single-dimension difference between two data objects (Bhat *et al.*, 2022). This metric reflects the greatest single-dimension difference between the objects, making it useful for scenarios where the maximum deviation in any one dimension is of particular interest.

- Another approach involves the *Cosine* distance,

$$d_{ij} = 1 - \frac{\sum_{n=1}^m x_{in} \cdot x_{jn}}{\left(\sum_{n=1}^m x_{in}^2\right) \left(\sum_{n=1}^m x_{jn}^2\right)}, \quad (3.9)$$

which measures the angular difference between two vectors rather than their absolute separation (Alfeilat *et al.*, 2019). The resulting fraction represents the cosine of the angle between the vectors and, subtracting it from one, gives the cosine distance, with values ranging from zero (indicating identical vectors) to two (indicating opposite vectors).

- For two sets J and S , the *Jaccard* distance,

$$d_{ij} = 1 - \frac{|J \cap S|}{|J \cup S|}, \quad (3.10)$$

is used to measure how dissimilar two sets are (Angkurawaranon *et al.*, 2022). Thus, a higher value of d_{ij} indicates greater dissimilarity between J and S .

- The *Hamming* distance,

$$d_{ij} = \sum_{n=1}^m |x_{in} - x_{jn}|, \quad (3.11)$$

calculates the difference between two data points by counting the number of positions where their corresponding elements differ (Bhat *et al.*, 2022).

The k-nearest neighbour algorithm is known for its simplicity and ease of implementation, but its performance can degrade in high-dimensional spaces (Defit *et al.*, 2023). Additionally, it can construct non-linear decision boundaries, allowing it to capture complex relationships in the data (Farid *et al.*, 2023). For these reasons, this algorithm is not widely employed in sentiment analysis (Kulkarni *et al.*, 2022). Increasing datasets and dimensions lead to more complex predictions, with all features receiving equal weight (Mladenova, 2021).

This algorithm relies on instance-based learning, making predictions based on the proximity of data points in a space. The Naïve Bayes classifier builds a statistical model during training and combines it with probability theory for predictions. This method is discussed next.

3.3.3.1.3 Naïve Bayes classifier

The Naïve Bayes classifier is a commonly used classification algorithm in the research community that relies on conditional probability calculations and makes predictions based on Bayes' theorem (Kulkarni *et al.*, 2022). The theorem assumes a conditional independence between the features, so each one contributes to the probability of a particular class, independently of the others (Bafjaish *et al.*, 2020). Therefore, it is also known as a probabilistic classifier method (Dara *et al.*, 2017). Bayes' theorem states that the probability of an event J given that event S has already occurred, $P(J|S)$ is expressed as

$$P(J|S) = \frac{P(S|J) \times P(J)}{P(S)}, \quad (3.12)$$

where $P(J)$ is the prior probability of the event J occurring and $P(S)$ is the marginal probability of the event S occurring (Jadhav *et al.*, 2024).

There are three types of naïve Bayes models, each suited for different types of data distributions, namely multinomial, Bernoulli, and Gaussian classifiers (Bafjaish *et al.*, 2020). The *multinomial* variant applies the Naïve Bayes method to input data that follows a multinomial distribution and counts the frequency of data (Quadri & Selvakumar, 2020). The *Bernoulli* algorithm is appropriate when the input data consists of binary values (Obaida & Salman, 2021). The *Gaussian* classifier is used in the case of continuous features that follow a normal distribution and estimates the likelihood of $P(J|S)$ based on the nature of the input data (Lakshmipriya *et al.*, 2023). Additionally, the classifier assumes that the input variables follow a Gaussian distribution for calculating their likelihood (Adhinata *et al.*, 2023; Obaida & Salman, 2021). Let $P(x_i|y)$ represent the probability density function of an input variable x_i given a class y , then

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \times \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right), \quad (3.13)$$

$$\text{with} \quad \mu_y = \frac{\sum_{i=1}^N x_i}{N}, \quad (3.14)$$

$$\text{and} \quad \sigma_y = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu_y)^2}{N}}, \quad (3.15)$$

where N is the number of samples in the dataset, x_i is the i^{th} sample, μ_y is the mean, σ_y^2 the variance, and σ_y is the standard deviation, given a class y . The normalisation factor in Equation 3.13 ensures that the total probability value adds to one, followed by the exponential function that determines how the probability decreases as x_i moves away from the mean.

Compared to other models, the Naïve Bayes classifier does not need much training time because it does not optimise iteratively, but calculates probability values (Johnson & Kelly, 2021). It is popular for its simplicity, effectiveness and easy implementation (Akhtar *et al.*, 2019; D'Andrea *et al.*, 2015). Although the estimates may occasionally be inaccurate, the approach is practical for

addressing multi-class feature problems and operates faster (Jaiswal & Singh, 2023). The classifier is considered computationally efficient because it operates under the simplistic assumption that data points are independent and generally performs well for text classification, particularly with small training datasets (Jadhav *et al.*, 2024). This, however, may not always hold true and can lead to suboptimal performance (Basuki, 2023). The assumption can cause the model to overlook complex relationships and correlations in the data, making it sensitive to the choice of variables (Hadinata *et al.*, 2020). Furthermore, Naïve Bayes may encounter a zero-frequency problem when a feature in the testing data does not appear in the training data, affecting the model's accuracy (Baboolal & Hosein, 2024).

3.3.3.1.4 Support vector machine

The support vector machine is a powerful non-probabilistic supervised learning technique frequently used for classification tasks (Kulkarni *et al.*, 2022). Support vector machines are a set of supervised learning methods used for classification, regression, and outlier detection, which are valuable for the statistical learning theory framework (Elbagir & Yang, 2019). These algorithms have shown strong accuracy in sentiment analysis and remain one of the most famous methods in the field (Akhtar *et al.*, 2019; Alam *et al.*, 2024). Furthermore, support vector machines are often implemented for text categorisation, hand-written digit recognition, tone recognition, image classification, and object detection (Bhambhu & Srivastava, 2010).

The four main constructs in support vector machines include hyperplanes, support vectors, margins, and kernel functions (Abd *et al.*, 2018). Consider Figure 3.6.

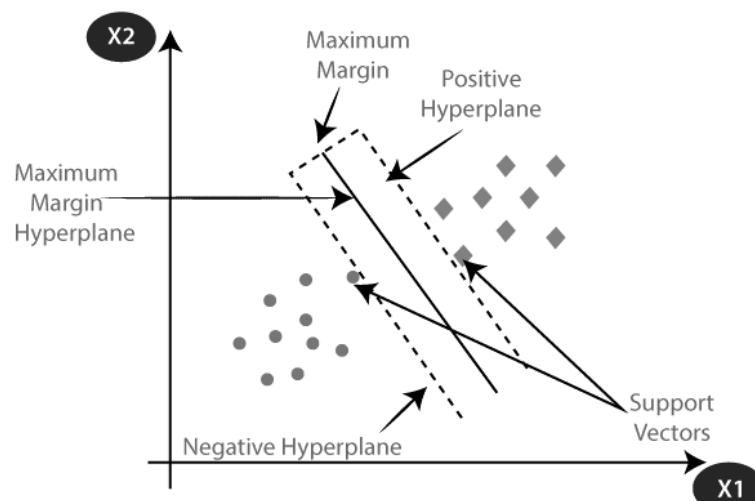


Figure 3.6: Visualisation of the support vector machine algorithm (Ghosh *et al.*, 2023)

In an n -dimensional space, a *hyperplane* is a flat affine subspace of dimension $n - 1$. *Support vectors* are the data points closest to the hyperplane. The machine is particularly focused on

these points because they are the most challenging to classify correctly and critical to defining the position and orientation of the hyperplane. The *margin* is defined as the distance between the hyperplane and the nearest data point from any class. Support vector machines seek to maximise this margin, creating the largest possible distance between the classes. A *kernel function* computes the scalar product of two vectors in a feature space, allowing the model to fit the maximum-margin hyperplane in a transformed feature space. The diagram illustrates two distinct groups of data points represented by circles and diamonds. A decision boundary separates these groups. The support vectors, which are the critical data points closest to the decision boundary, are shown as either circles or diamonds. These support vectors lie on two parallel hyperplanes: the negative hyperplane, which is closer to the circles and the positive hyperplane, which is closer to the diamonds. The maximum margin hyperplane is the optimal decision boundary that lies equidistant between the two parallel hyperplanes. The maximum margin refers to the distance between the positive and negative hyperplanes, which the support vector machine maximises to ensure the best possible separation between the two groups. This margin is crucial for the classifier's ability to generalise well to unseen data. The algorithm employs hyperplanes in a high-dimensional space to analyse data and establish decision boundaries (Kulkarni *et al.*, 2022; Sharma & Tyagi, 2018). Hyperplanes are also referred to as boundaries that distinguish between clusters of data (Hofmeyr, 2023).

Data classification using support vector machines is broadly divided into two types, linear classification using the maximum margin hyperplane and non-linear classification using kernel functions (Ahamed *et al.*, 2022). Common kernel functions include polynomial, radial basis, exponential and sigmoid functions (Bhambhu & Srivastava, 2010). A hybrid kernel function can also be used by combining multiple basic kernel functions (Jun, 2021).

The *polynomial* kernel function measures the similarity between two vectors in a transformed feature space as

$$K(a, b) = (a^T b)^q, \quad (3.16)$$

where, a and b are the input vectors, $a^T b$ represents the scalar product of these vectors, and q is a positive integer that determines the degree of the polynomial (Jun, 2021). The dot product operation computes the sum of the products of corresponding elements of a and b . Raising the scalar product to the power of q effectively maps the original vectors into a higher-dimensional space, allowing the algorithm to capture more complex relationships in the data. The polynomial kernel thus enables the learning algorithm to create non-linear decision boundaries, making it more potent for handling complex datasets. When $q = 1$, the polynomial kernel reverts to the linear kernel.

Another function that measures the similarity between two data points is the *radial basis* function, also known as the Gaussian similarity kernel. This is calculated as

$$K(a, b) = \exp\left(-\frac{(a-b)^2}{2\sigma^2}\right), \quad (3.17)$$

where $(a - b)^2$ represents the squared Euclidean distance between the points a and b , and σ is a scaling factor (Min *et al.*, 2018). The scaling factor determines the width of the Gaussian function, controlling the spread of the kernel. The denominator $2\sigma^2$ normalises the squared distance, and the exponential function ensures that the result lies between 0 and 1. When the points a and b are close, the value of $K(a, b)$ approaches 1 which indicates a high similarity. When they are far in the distance, the value of $K(a, b)$ approaches 0 which indicates a low similarity. This kernel function is crucial for mapping input data into a higher-dimensional space, where tasks such as classification or regression become easier to perform.

The *exponential* kernel function is a type of radial basis function for handling non-linear relationships and is calculated as

$$K(a, b) = \exp\left(-\frac{|a-b|}{\sigma}\right), \quad (3.18)$$

where $\exp$ is the base of the natural logarithm e raised to a power to be computed, $|a - b|$ represents the absolute difference between the points a and b , σ is a positive scaling factor (Jun, 2021). This kernel is commonly used in machine learning and statistical tasks such as clustering and classification because it helps to assess how closely related two data points are based on their distance.

The *sigmoid* kernel function, also known as the hyperbolic tangent similarity function, is calculated as

$$K(a, b) = \tanh(ca^T b + h), \quad (3.19)$$

where c is the scaling factor that controls the slope of the function, h is a bias term, and a^T denotes the transposition of the vector a , allowing for the scalar product $a^T b$ (Min *et al.*, 2018). The hyperbolic tangent function, $\tanh$, gradually transitions from zero and increases as the data points become more aligned in direction, achieving higher similarity values for larger inputs.

Although support vector machines perform very well in high-dimensional spaces, they are not as effective on large datasets (Jaiswal & Singh, 2023). It takes a long time to train on large datasets, and model fine-tuning is time-consuming and challenging (Du *et al.*, 2023).

3.3.3.1.5 Logistic regression

Logistic regression is a form of regression analysis utilised in classification tasks, predicting the probability that a given input belongs to a particular class (D'Andrea *et al.*, 2015; Kulkarni *et al.*, 2022). The logistic, or sigmoid function, plotted in Figure 3.7, is a component of the logistic

regression formula that ensures predictions fall within the range of 0 or 1 (Arief & Samsudin, 2023).

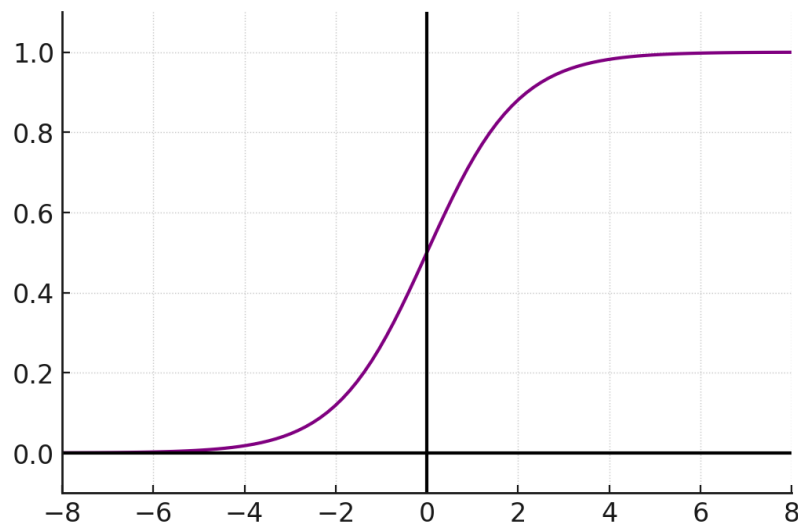


Figure 3.7: Visualisation of the sigmoid function (Deshpande *et al.*, 2023)

A sigmoid function forms an S-shaped curve and is commonly used in machine learning and statistical models, particularly in logistic regression and neural networks. The curve gradually increases from 0, passes through 0.5 at the midpoint, and asymptotically approaches 1 as the input value increases. This smooth transition allows the sigmoid function to effectively classify data points and model binary outcomes where the output represents the probability of a certain class or event occurring.

The predicted probability is converted into a categorical output by applying a cutoff threshold (Reddy *et al.*, 2021). The optimal cutoff can be selected based on the specific business case, for example, balancing precision and recall or minimising the cost of false positives and false negatives. Observations with a probability higher than the cutoff are classified as one class, while those with a probability lower than the cutoff are classified as the other class (Geisler *et al.*, 2022). The model is known for quickly classifying unidentified data, but it is prone to overfitting (Jaiswal & Singh, 2023). Moreover, boundaries are linear and unable to handle complex non-linear problems, resulting in low accuracy with complicated datasets (D'Andrea *et al.*, 2015). Despite these limitations, the model was selected for use in the current study due to its simplicity, interpretability, and computational efficiency, making it a suitable baseline for comparison. To mitigate issues such as overfitting and linearity constraints, hyperparameter optimisation was applied, and the model's performance was evaluated alongside more advanced algorithms to ensure robustness.

The main types of logistic regression are binomial or binary logistic and multinomial logistic regression (Brito-Loeza *et al.*, 2020; Ghous *et al.*, 2021). Binary logistic regression addresses situations where the dependent variable can only result in two possible outcomes. Often referred to as maximum entropy modelling (Dara *et al.*, 2017), it is mathematically expressed as:

$$\text{logit}(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X, \quad (3.20)$$

$$\text{with } \pi = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}, \quad (3.21)$$

where Y is the target variable, π is the probability of an event occurring, α is the Y -intercept, β the regression coefficient, and X is the input or independent variable (Ingersoll *et al.*, 2002). The coefficient β indicates the direction of the relationship between X and the *logit* of Y : a positive β means that higher X values correspond to higher *logits* of Y while a negative β means that higher X values correspond to lower *logits* of Y . Multinomial logistic regression is an extension of binary logistic regression and is often referred to as *softmax* regression (Elbagir & Yang, 2019). However, it addresses situations where the dependent variable can result in three or more possible outcomes, which was not considered for this study.

Binary logistic regression will be used in this study to model sentiment classification, where the dependent variable represents two possible outcomes. Several assumptions accompany the use of this model (Pal, 2021):

- Linearity of the *logit*: The relationship between the independent variables and the logarithm of the odds of the dependent variable, is linear.
- Independence of errors: Observations should be independent of one another, meaning the outcome of one observation does not influence the outcome of another.
- Absence of multicollinearity: Independent variables should not highly correlate with each other to avoid redundancy and ensure the stability of coefficient estimates.
- Large sample size: Logistic regression requires a sufficiently large sample size to produce reliable and stable estimates, especially when dealing with rare events.
- Binary outcome for binary logistic regression: The dependent variable should be binary, meaning it has only two possible outcomes.
- Appropriate link function: The logistic regression model assumes the use of the logistic (sigmoid) function as the link function, which maps predicted values to probabilities between 0 and 1.

To gain insight into model behaviour and generalisation capability, performance metrics such as training and validation loss are evaluated (Cordova-Esparza *et al.*, 2025). These metrics provide an indication of how well the model learns from the training data while maintaining performance on unseen data, thereby assisting in identifying potential issues such as underfitting or overfitting

(Bezemer *et al.*, 2024). Loss is typically defined using the logarithmic loss function, which measures the divergence between the predicted probabilities and the actual class labels. The logarithmic loss function measures the performance of a classification model by comparing the predicted probability p with the true class label y and is mathematically expressed as

$$L(y, p) = -(y \log(p) + (1 - y) \log(1 - p)), \quad (3.22)$$

where the first term, $y \log(p)$, penalises the model when it assigns a low probability to the true class, while the second term, $(1 - y) \log(1 - p)$, penalises incorrect predictions for the class. The negative sign ensures that the loss value is positive, with smaller values indicating better model performance (Cordova-Esparza *et al.*, 2023). This loss function is susceptible to the confidence of probabilistic predictions, as highly confident but incorrect predictions incur a large penalty. Monitoring both training and validation loss allows detection of issues such as overfitting or underfitting (Prajapati & Singh, 2023). If the training loss is significantly lower than the validation loss, it is an indication that a model fits the training data too well, or overfits, but struggles to generalise to unseen data. Conversely, high loss for both datasets indicates underfitting, where the model is too simple to capture the underlying patterns in the data.

The simplicity, effectiveness, and interpretability of logistic regression make it a favourable choice when it comes to developing a predictive model (Amirkumar *et al.*, 2024). It is the simplest model for a classification task, and therefore, it serves as an appropriate baseline model for this study.

3.3.3.1.6 Random forest

Also known as random decision forests, this is a machine-learning method that builds multiple decision trees to classify, regress, or predict outcomes by aggregating the results from individual trees (Ghous *et al.*, 2021). Jaiswal and Singh (2023) added that although it seems more complicated to use, such a robust model performs accurately. It is suitable for regression (where the target variable is numeric) and classification (where the target variable is categorical) modelling, excelling in classification tasks (Deshpande *et al.*, 2023; Elbagir & Yang, 2019).

Random forests are widely regarded as one of the most effective classification techniques in machine learning (Ahamed *et al.*, 2022). It addresses the problem of overfitting by averaging the outcomes of multiple decision trees (Kanozia *et al.*, 2023). Different splits and divisions will result from each tree in the forest because of varying data subsets and attributes, leading to a different classification (Jaiswal & Singh, 2023). It is, therefore, an ensemble of decision trees that enhances prediction accuracy and stability (Amirkumar *et al.*, 2024).

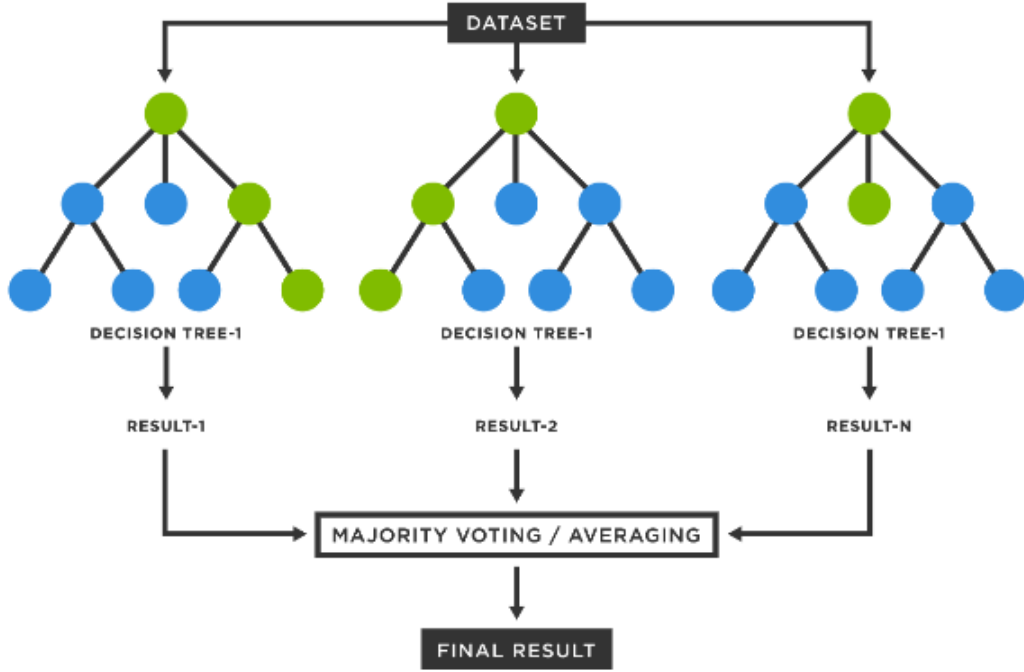


Figure 3.8: The random forest algorithm (Deshpande *et al.*, 2023)

Figure 3.8 illustrates a case where multiple decision trees are used to improve the accuracy and robustness of the final prediction. Budi *et al.* (2020) summarised the process as follows:

- *Bootstrapping* involves the creation of a random data sample of size n from the dataset to facilitate group recovery.
- A random feature selection is performed by constructing a decision tree to its maximum size without pruning. This involves randomly choosing an explanatory variable, ensuring that the number of selected variables, m , is less than the total number of available variables, p . This procedure is repeated n times, resulting in a forest with n trees.

By repeating the process, a random forest is created (Cao *et al.*, 2021). It is mathematically expressed as

$$P(y = c|x) = \frac{1}{N} \times \sum_{n=1}^N (Tree_n(x) = c), \quad (3.23)$$

where $P(y = c|x)$ denotes the probability that the output y equals the class c given the input x , with N the total number of decision trees in the ensemble, and $Tree_n$ is the n^{th} tree (Kanozia *et al.*, 2023). By averaging the sum of all trees classifying the input x as class c , the function yields the probability that the ensemble classifies x as class c . This approach leverages the collective decision-making of multiple trees to provide a more robust and accurate classification probability.

Shallow learning models are known for their simplicity, efficiency, and interpretability. However, some tasks require models that can automatically learn complex patterns from large amounts of data.

3.3.3.2 Deep learning

A deep learning approach contains multiple hidden layers, or transformations, between the input and output layers. It is a class of machine learning inspired by the human brain's biological neuron structure and function (Abaye *et al.*, 2020). Drawing an analogy to the human brain, deep learning algorithms are typically implemented as artificial neural networks.

A neural network is a parallel computing system composed of numerous simple processors that are mathematically interconnected (Kannan & Ramachandran, 2021). Such an artificial model can automatically learn hierarchical representations from raw data, making them particularly effective for complex tasks such as image and speech recognition, NLP, sentiment analysis, and more. A typical network consists of anywhere from a few to millions of processing elements called artificial neurons, also known as units or nodes. Figure 3.9 illustrates the difference between a shallow neural network and a deep neural network.

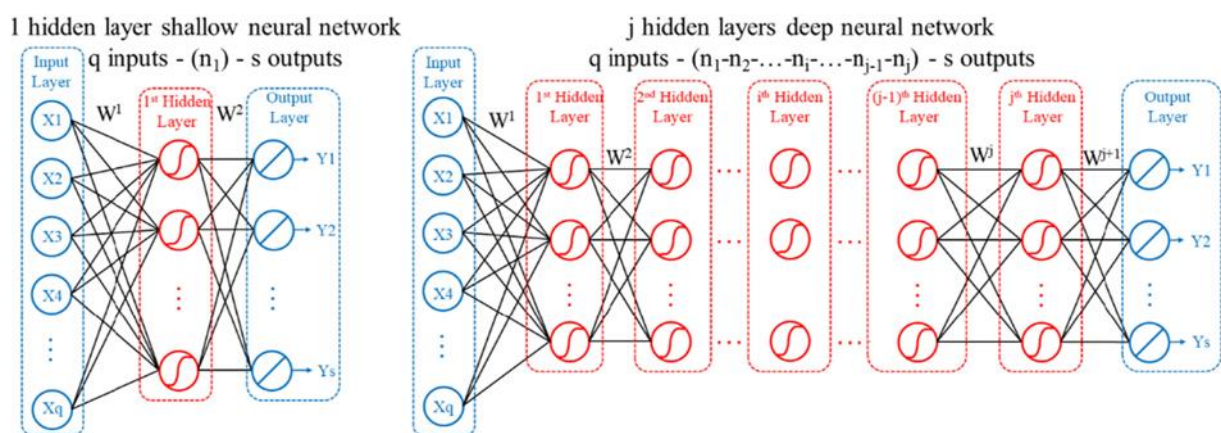


Figure 3.9: A single hidden layer shallow neural network vs a j -hidden layer deep neural network (Dong *et al.*, 2019)

Neurons in a deep neural network are interconnected building blocks that process and transmit information within a network. Each neuron applies a mathematical function like a weighted sum of its inputs, adds a bias term, and passes the result through an activation function to introduce non-linearity. As data flows through successive layers, these transformations allow the network to learn increasingly abstract and high-level features from raw inputs.

However, deep learning approaches require significantly more computational power, larger datasets, and careful adjustment of hyperparameters, or tuning, to ensure accurate prediction results (Aloqaily *et al.*, 2020). Hyperparameters may include the number of layers, the number of neurons per layer, the activation functions, the learning rate, batch size, and many more (Shaziya & Zaheer, 2021). Hyperparameters will be discussed in more detail in the Chapter 6.

The increased depth provided by many layers allows for capturing intricate relationships but also makes model training more challenging due to issues like vanishing gradients. From theoretical studies, it can be deduced that deep networks outperform shallow networks in prediction accuracy due to their hierarchical structure and their ability to represent complex functions with fewer parameters (Bassett *et al.*, 2020; Liao *et al.*, 2017; Liao & Poggio, 2018; Shapira, 2023). Consider the simplified deep neural network in Figure 3.10.

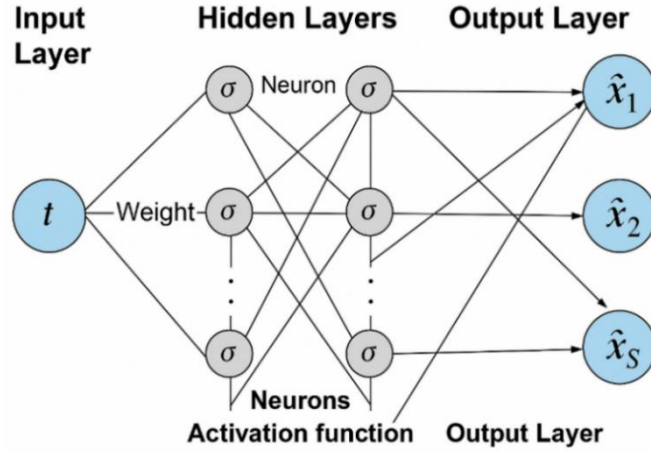


Figure 3.10: An example of a deep neural network

Each connection between layers carries a weight, while each neuron is associated with a bias (threshold) and an activation function that determines its output (Balakrishnan *et al.*, 2019). To minimise the error between the predicted and actual output values, the network updates its weights and biases through two key processes: forward propagation and backpropagation (Al-Alawi, 2002; Cheng *et al.*, 2012; Meghana & Sekhar, 2020; Nath, 2024).

Forward propagation is the process of feeding input data into the neural network and processing it through its layers to generate predictions. The complete algorithmic formulation is provided in Algorithm 3.1.

Algorithm 3.1: Forward propagation in a neural network

Input: X_1 is the input layer vector (function input);
 $[W_{12}, W_{23}, \dots, W_{m-1,m}]$: the weights list;
 $[b^{(2)}, b^{(3)}, \dots, b^{(m)}]$: the bias list;
 $g(\cdot)$ is the activation function.

Output: X_m is the network output.

```

1   $X \leftarrow X_1$ ;
2   $layer\_list \leftarrow [X_1]$ ;
3  for  $j$  in range (2 to  $m$ ) do
4      //linear combination (pre-activation)
5       $u^{(j)} \leftarrow W_{j-1,j} \cdot X + b^{(j)}$ ;
6      //activation
7       $X \leftarrow g(u^{(j)})$ ;
8      append  $X$  to  $layer\_list$ ;
```

```

9   end for
10   $X_m \leftarrow X$ ;
11  output  $X_m$ ;

```

During forward propagation, the output of each neuron is computed by applying an activation function to the weighted sum of its inputs and the associated bias term. This computation proceeds sequentially from the input layer through the hidden layers to the output layer, producing the final predicted value. Forward propagation, therefore, represents the process by which input features are transformed into model predictions through successive linear combinations and non-linear activations. Once the predicted output is obtained, a loss function is used to quantify the difference between the predicted and actual target values.

In backpropagation, the gradients of this loss for each weight are then calculated by propagating the error backwards through the network. Using these gradients, an optimisation algorithm such as gradient descent, adjusts the weights to reduce the error, gradually refining its parameters to improve performance on the training data. The algorithm is formally presented in Algorithm 3.2.

Algorithm 3.2: Backpropagation in a neural network

Input: Training example (x_i, y_i) ;
 $[X_1, X_2, \dots, X_m]$: the layer list;
 $[u^{(2)}, u^{(3)}, \dots, u^{(m)}]$: the pre-activations;
 $[W_{12}, W_{23}, \dots, W_{m-1,m}]$: the weights list;
loss $l(y_i, \hat{y}_i)$;
 $g'(\cdot)$ is the derivative of the activation function;
 η is the learning rate.

Output: Gradients $\nabla_{\theta} l$ and updated parameters θ .

```

1  //Base case
2  initialise  $\frac{\partial l}{\partial l} \leftarrow 1$ ;
3  //Assume forward pass already computed
4  //Compute node-wise gradients using chain rule from output layer backwards
5  for each node  $u$  in a given layer do // from the last layer to the first hidden layer
6      for each node  $v$  in  $\Gamma(u)$  do //  $\Gamma(u)$  is the nodes in the next layer
7          //Use already computed  $\frac{\partial l}{\partial v}$  to calculate the following
8          compute  $\frac{\partial v}{\partial u} | x_i, y_i, \theta$ ;
9          accumulate  $\frac{\partial l}{\partial u} \leftarrow \sum (\frac{\partial l}{\partial v} \cdot \frac{\partial v}{\partial u})$ ;
10     end for
11     //Convert node gradients to parameter gradient
12     compute  $\frac{\partial l}{\partial W} \leftarrow \frac{\partial l}{\partial u} \cdot \frac{\partial u}{\partial W}$  //  $\frac{\partial u}{\partial W}$  depends on previous-layer values;
13     compute  $\frac{\partial l}{\partial b} \leftarrow \frac{\partial l}{\partial u} \cdot \frac{\partial u}{\partial b}$ ;
14 end for
15 //Gradient descent parameter update
16 update  $\theta \leftarrow \theta - \eta \nabla_{\theta} l$ ;
17 output updated  $\theta$ ;

```

The purpose of an activation function is to introduce non-linearity into the network, enabling it to learn complex patterns and to ensure stable model convergence (Chen *et al.*, 2019). Commonly used activation functions are summarised in Table 3.1.

Table 3.1: Activation functions used in neural networks (Athaiya *et al.*, 2020; Laxkar & Tomar, 2022)

Function	Output range	Applications	Advantages	Disadvantages	Mathematical form
Binary step	$\{0,1\}$	Simple binary classification.	Simple, fast for binary classification.	Non-differentiable, leads to poor training with gradient-based methods.	$binary\ step(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$
Linear	$(-\infty, \infty)$	Regression tasks.	Simple, good for regression tasks.	Can't model non-linearity, may lead to poor performance in complex problems.	$Linear(x) = x$
ReLU (Rectified Linear Unit)	$[0, \infty)$	Most deep networks (CNNs, feedforward)	Efficient computation, helps with sparse activations.	Neurons can output zero permanently (die).	$ReLU(x) = \max(0, x)$
Leaky ReLU	$(-\infty, \infty)$	CNNs, where ReLU causes dead neurons.	Prevents dying neurons, still efficient computation	Small negative slope might not fix a dying ReLU.	$Leaky\ ReLU(x) = \begin{cases} x, & x \geq 0 \\ \alpha x, & x < 0 \end{cases}$ with α a positive constant.
ELU (Exponential Linear Unit)	$(-\alpha, \infty)$	Deep networks requiring faster convergence,	Speeds up convergence, produces smoother output.	Slightly more computationally expensive.	$ELU(x) = \begin{cases} x, & x > 0 \\ \alpha(e^x - 1), & x \leq 0 \end{cases}$ with α a positive constant.
PReLU (Parametric ReLU)	$(-\infty, \infty)$	Deep networks where Leaky ReLU is needed but with a learned slope.	Learns slope for negative inputs, potentially improves performance.	Increased complexity due to learning an extra parameter.	$PReLU(x) = \max(\alpha x, x)$ where α is learned.
Sigmoid	$(0,1)$	Binary application, early neural networks.	Smooth gradient, works well for binary classification.	Vanishing gradients, non-zero-centred output.	$\sigma(x) = \frac{1}{1 + e^{-x}}$
Tanh	$(-1,1)$	RNNs, networks requiring zero-centred activation.	Zero-centred, helps with symmetry of data.	Suffers from vanishing gradients.	$tanh(x) = \frac{e^x - e^{-1}}{e^x + e^{-x}}$

Swish	$(-0.28, \infty)$	Deep networks, alternative to ReLU	Smooth, continuous function, works well in deep networks	Newer, more complex, and not always better.	$\text{Swish}(x) = x \cdot \sigma(x)$ $= \frac{x}{1 + e^{-x}}$
-------	-------------------	------------------------------------	--	---	--

Activation functions introduce non-linearity to neural networks which enable them to model complex relationships in data. Choosing the appropriate one can notably impact the performance and efficiency of such a network. With the introductory concepts of deep learning established, some of the different architectures will be discussed.

3.3.3.2.1 Convolutional neural network

The CNN was first introduced in the 1960s and has become a leading deep learning algorithm that excels in complex visual tasks such as image classification, segmentation, and object detection and extends to video processing, NLP, and speech recognition (Fathurrahman *et al.*, 2023). The networks are known as feedforward neural networks, as information flows in one direction, from input to output (Deveci *et al.*, 2024).

Data can undergo several transformations between the input and output stages. Consider the diagrammatic representation of a CNN in Figure 3.11.

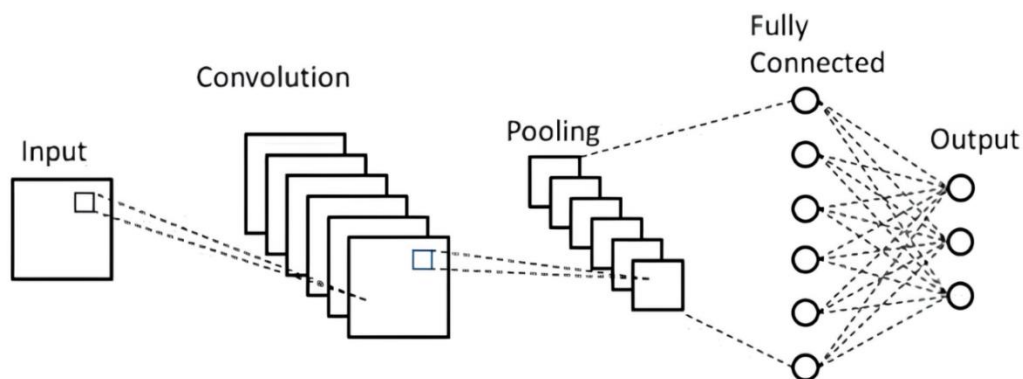


Figure 3.11: Architecture of a CNN (Phung & Rhee, 2019)

The operation of a CNN starts at the **input layer**, where raw data, such as images, is fed into the network. The input data is filtered in the convolutional layer to detect local patterns like edges or textures, resulting in feature maps. These maps are then passed through the pooling layer, which reduces their dimensionality by retaining only the most significant information, increasing model efficiency. After pooling, the data is flattened and sent to the fully connected layer, where high-level reasoning occurs by integrating the features learned in previous layers. Finally, the output layer generates the final prediction which could be a classification label or another relevant output

depending on the task. This structured approach enables CNNs to automatically learn and extract features from images which makes them highly effective for complex visual tasks.

The **convolutional layer** is the most essential part of the structure and consists of a collection of kernel filters that are applied to the data. A kernel is a set of discrete values or integers used to detect specific features in the data (Taye, 2023). Each kernel is represented by neurons that form a filter with three dimensions, namely length, width, and depth (Deveci *et al.*, 2024; Fanani *et al.*, 2021). The convolutional layer is mathematically written as

$$h_{w,b}(x) = f(W^T x) = f(\sum_{i=1}^n W_i \cdot x_i + b), \quad (3.24)$$

where W denotes the weight vector that determines the importance of each input feature in the final output, x is the input vector, b is the bias term, and f is an activation function (Cui, 2018). The term $W^T x$ denotes the dot product between the weight vector and the input vector, which produces a weighted sum of the inputs. The bias is added to adjust the output. Finally, the activation function is applied to this weighted sum to introduce non-linearity into the model, enabling it to capture complex relationships in the data. This equation is fundamental in neural networks, where f can be any activation function, depending on the desired characteristics of the model. Figure 3.12 illustrates the operation of a single computing unit in a convolutional network.

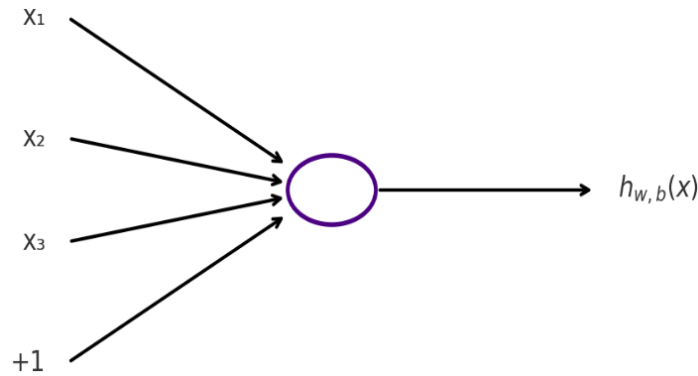


Figure 3.12: The function of a single computing unit in a convolutional layer (Cui, 2018)

A basic computational unit, often referred to as a perceptron, is a foundational concept in neural networks. In this diagram, the inputs x_1 , x_2 and x_3 represent the features fed into the model. Each input is associated with a corresponding weight that determines its contribution to the final output. The additional input labelled $+1$ corresponds to the bias term, which helps shift the output to improve model accuracy. These inputs are combined in the computational unit (represented by the circle), which is linearly weighted and summed. This sum is then passed through an activation function, resulting in the output, $h_{w,b}(x)$. The illustrated model learns appropriate weights and bias during training to make accurate predictions.

The **pooling layer**, also known as the subsample layer, is primarily used to decrease the spatial dimensions of the input data, extract features, and reduce computation loads by combining two consecutive convolutional layers (Du *et al.*, 2018; Fathurrahman *et al.*, 2023). Moreover, it minimises overfitting by decreasing the number of parameters, which also lowers the overall cost⁶ (Ayeni, 2022). Commonly used pooling functions include the *max* and *average* pooling functions (Deveci *et al.*, 2024). Like the word says, max pooling selects the largest element within each pooling region, according to

$$y_{kij} = \max_{(p,q) \in R_{ij}} x_{kpq}, \quad (3.25)$$

where y_{kij} is the output of the maximum pooling operation at position (i, j) in channel k , x_{kpq} is the input value at position (p, q) in channel k , and R_{ij} is the pooling region corresponding to the output (i, j) (Jie & Wanda, 2020). Essentially, max pooling reduces the spatial dimensions of the input by summarising each region with its maximum value, thus capturing the most prominent features in that area. *Average* pooling describes the calculation of the arithmetic mean of the values within each pooling region and is formulated as

$$y_{kij} = \frac{1}{|R_{ij}|} \sum_{(p,q) \in R_{ij}} x_{kpq}, \quad (3.26)$$

with y_{kij} the output of the average pooling operation at position (i, j) in channel k , x_{kpq} is the input value at position (p, q) in channel k , R_{ij} is the set of indices in the pooling region corresponding to output (i, j) , and $|R_{ij}|$ the number of elements in the pooling region (Jie & Wanda, 2020). The averaging operation helps reduce the dimensionality of the feature map while retaining important spatial information. Consider the operation of the two methods displayed in Figure 3.13.

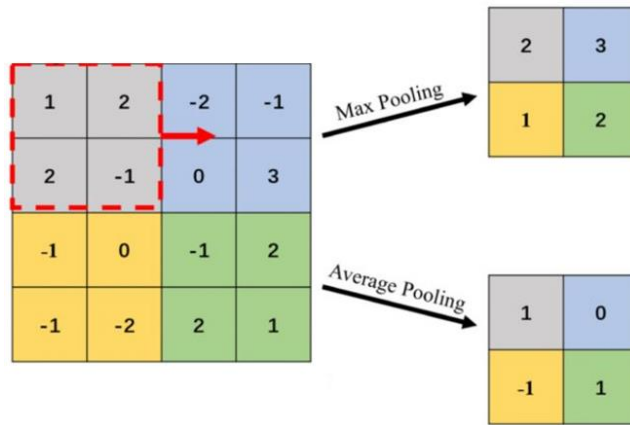


Figure 3.13: The max vs average pooling approaches (Branko *et al.*, 2022)

⁶Cost refers to the computational cost of training the model, which includes the time and resources required for processing, memory usage, and the number of operations needed to optimise the network.

A 4×4 input matrix is divided into four 2×2 regions, each represented by different colours. In max pooling, the maximum value from each 2×2 region is extracted. These maximum values are then combined to form a new 2×2 matrix, as depicted in the top-right portion of the image. The average pooling function will calculate the average value for each 2×2 region and create a different 2×2 matrix, shown in the bottom-right section of the image. Both max pooling and average pooling serve to reduce the dimensions of the input matrix but do so in different ways. Pooling methods help decrease computational complexity and prevent overfitting in neural networks, making them components of CNN architectures.

A **fully connected layer** performs a global operation, unlike convolution and pooling layers, and is usually used at the end of the network for classification purposes (Deveci *et al.*, 2024). Output from the final convolution or pooling layer is flattened into a one-dimensional array and fed into fully-connected layers, or dense layers, where each input is linked to every output node through learnable weights (Do *et al.*, 2018). Gradient descent, or batch mode learning, minimises the cost function by evaluating the entire dataset and updating parameters only after completing one full pass through the data (Asopa *et al.*, 2018). It iteratively/repeatedly reduces any training error through parameter updates guided by the gradient of the objective function. These updates are applied in the opposite direction of the gradient by means of back-propagation (Al-Amidie *et al.*, 2021). Neurons are grouped into layers, with each node fully connected to every node in the adjacent layers. Nodes in the pooling layer are flattened into vectors and input to the fully-connected layer, which ultimately connects to the output layer. These last two layers are illustrated in Figure 3.14.

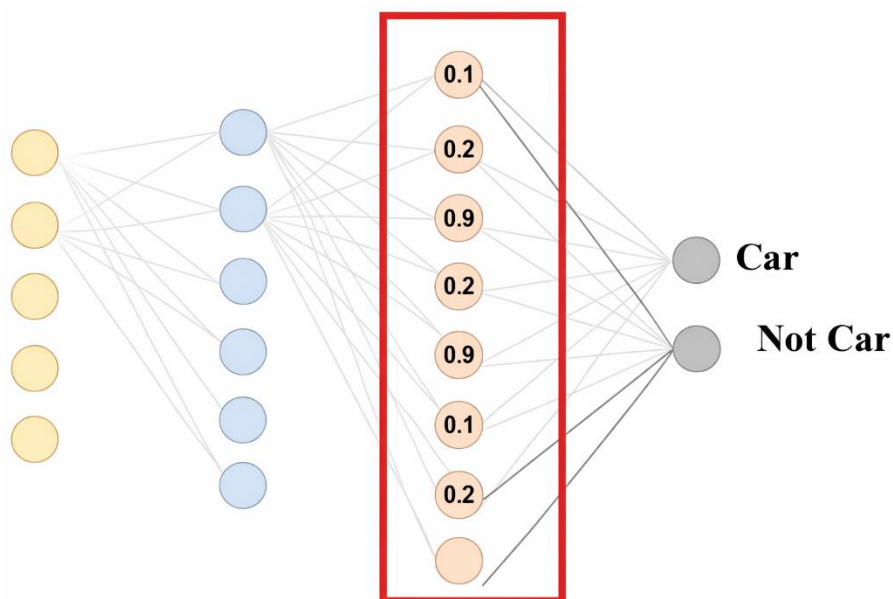


Figure 3.14: Example of a fully connected layer linked to an output layer for the identification of a car (Taye, 2023)

The goal of the CNN, whose final layers are depicted in the picture, is to determine whether an input image represents a car. The network comprises multiple layers, of which the fully-connected and output layers are illustrated. A layer where each node outputs a probability score between 0 and 1 is highlighted in the red box. These scores represent the confidence level of the network for each feature in the layer. The nodes with higher scores (e.g. 0.9) indicate more substantial confidence in the presence of specific features that contribute to the classification decision.

These probability scores are then supplied to the final **output layer**, which consists of two nodes corresponding to the possible classifications. The network ultimately selects the class with the higher aggregated score from the previous layer. This image effectively demonstrates how a neural network processes input features through layers of abstraction to arrive at a classification decision, with the highlighted layer playing a crucial role in determining the outcome. The output layer, often seen as the mathematical layer, is responsible for interpreting the numerical patterns that have been identified in preceding tasks (Ayeini, 2022).

CNNs are generally easier to train than other types of networks because they have fewer parameters. This reduces the risk of overfitting and efficiently combines feature extraction with classification, making them more practical for implementation in large networks (Asopa *et al.*, 2018). However, they automatically learn spatial feature hierarchies through back-propagation and are not ideal for applications requiring stock price prediction that relates to sequential data (Do *et al.*, 2018).

3.3.3.2.2 Recurrent neural network

Recurrent neural networks (RNNs) are specifically designed to process sequential data, where the order of the inputs is critical (Mienye *et al.*, 2024). Unlike feedforward neural networks, RNNs have connections forming cycles that allow information to persist over time (Ji *et al.*, 2020). This makes them particularly effective for tasks involving time-series data like speech processing, NLP, and other sequential data applications (Ackerson *et al.*, 2021; Al-Amidie *et al.*, 2021). Furthermore, RNNs can maintain a hidden state that captures information from previous time steps (Lu & Sun, 2022).

The hidden state is updated at each timestep based on the current input and the previous hidden state (Lim, 2021). A hidden step is maintained according to

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \quad (3.27)$$

$$y_t = \sigma(W_{hy}h_t + b_y), \quad (3.28)$$

where h_t is the state at time step t , x_t the current input, with h_{t-1} the previous hidden state, b_h a bias term, W_{xh} and W_{hh} weight matrices, σ is an activation function (commonly the sigmoid or

$\tanh$), and y_t the final output (Addepalli *et al.*, 2020). At each time step, the hidden state is updated by passing a weighted sum of the input and previous state through an activation function. The output is then calculated by applying a weight matrix to the hidden state and adding a bias term, which is further processed through another activation function. The typical architecture of an RNN model is illustrated in Figure 3.15.

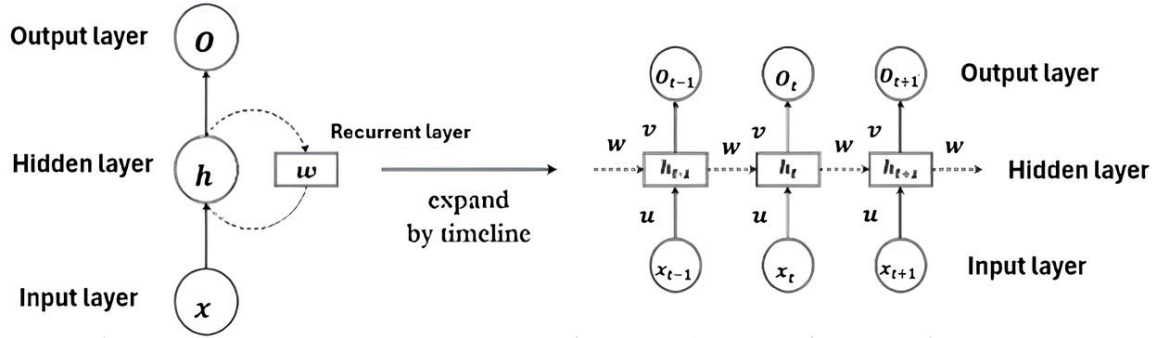


Figure 3.15: Architecture of an RNN (Wu, 2022)

The basic structure of an RNN is shown on the left, along with the input, hidden and output layers. The hidden layer plays a crucial role, as it not only receives the current input but also maintains a connection to its previous state through the recurrent weight, w . This enables the network to preserve a memory of prior inputs and incorporate this historical context when processing new data points. The right part of the diagram shows how the network progresses over time, demonstrating how the model handles sequential data by retaining hidden states across time steps. Each time step has its own input, hidden state and output.

This recurrent structure enables the network to process information in a time-dependent manner, making RNNs suitable for tasks such as time series forecasting, language modelling, and speech recognition. By expanding along the timeline, the RNN can take past information into account, giving it an advantage over traditional feedforward neural networks for tasks involving sequences. While RNNs can theoretically handle time-series data, their limitations in capturing long-term dependencies, handling noise, and managing complex relationships make them less ideal for problems like stock prediction. Accordingly, Long Short-Term Memory (LSTM) networks were developed by incorporating mechanisms that allow them to capture and retain long-term dependencies.

3.3.3.2.2.1 Long Short-Term Memory model

The LSTM model is a type of RNN specifically designed to learn from sequential text data (Anumala *et al.*, 2019; Fuller, 2022; Gupta *et al.*, 2021; Othlasseril & Shaikh, 2021). It has a "no-forget" feature that, for example, can remember past stock prices, which is crucial when predicting

future stock prices (Othalasseril & Shaikh, 2021). The LSTM can effectively capture and learn long-term dependencies and extract features that other models fail to do, distinguishing it from basic RNNs (Anumala *et al.*, 2019; Kotecha *et al.*, 2021; Nandwani & Verma, 2021). Furthermore, the context of a sentence is considered, rather than just a single word, as well as part of a sentence, and the word sequencing in a sentence (Fuller, 2022). This model has proven to be one of the most successful when applied to sentiment data (Gupta *et al.*, 2021). Explaining the operation of the LSTM network is based on the diagram in Figure 3.16.

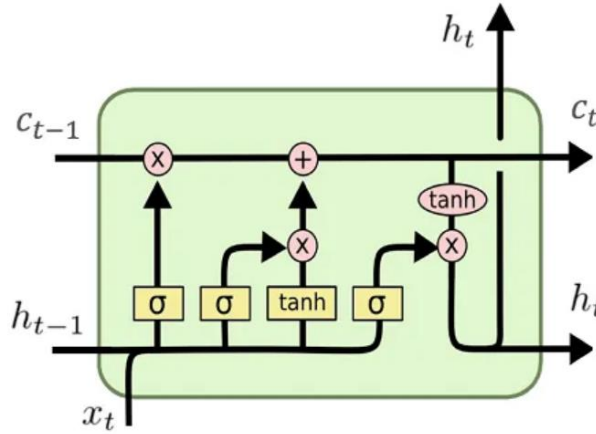


Figure 3.16: The operation of an LSTM operation (Rahuljha, 2020)

The operational framework of the LSTM network revolves around how it processes and retains information over time using its *memory cell* and *gating mechanisms* (Gao *et al.*, 2022; Li & Zhang, 2023). Data is input one step at a time, while updating the internal memory is based on the current input and past states.

The cell state resembles a continuous belt running through the entire sequence and can be updated, deleted or added to as needed by the gates. The three gates controlling the flow of information are the *forget gate* that decides what information to discard, the *input gate* that determines which information to store, and the *output gate* that controls which information to output (either as a prediction or for the next iteration). These gates are represented mathematically as:

$$\text{(input gate)} \quad i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \quad (3.29)$$

$$\text{(forget gate)} \quad f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \quad (3.30)$$

$$\text{(memory cell)} \quad c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (3.31)$$

$$\text{(output gate)} \quad o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o), \quad (3.32)$$

$$\text{(new hidden state)} \quad h_t = \sigma_t \tanh(c_t), \quad (3.33)$$

where σ denotes the sigmoidal activation function, W_i represents the weight matrices with $i \in \{\text{input, forget, output, memory}\}$. The sigmoid function is applied to individual elements. The weight matrices are essential in linking the input to components like the input gate (i_t), forget gate

(f_t) , output gate (o_t) , and memory cells (c_t) . Here, x_t denotes the input vector at time step t , h_{t-1} is the hidden state from the previous time step, b_i , b_f , b_c and b_o are the bias terms for each respective gate, and t represents the current time step in the sequence.

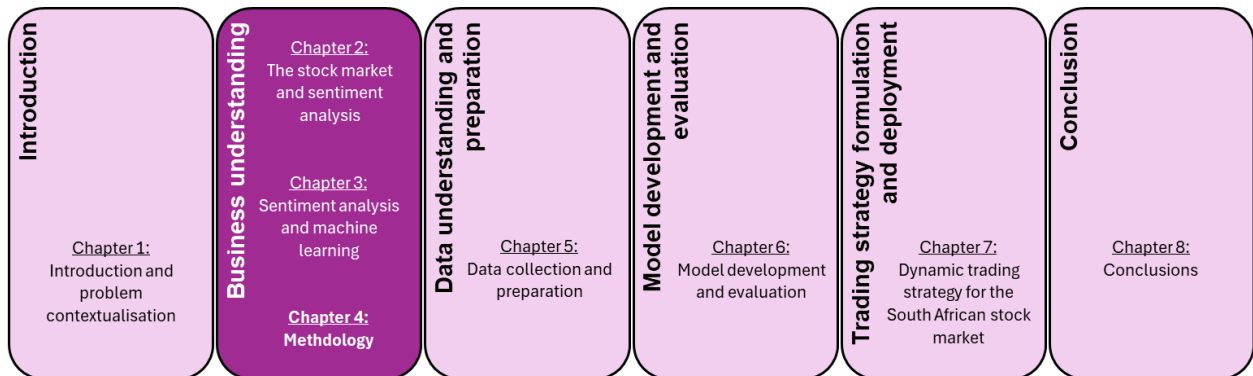
The strength of the LSTM network lies in its ability to selectively remember or forget information, making it ideal for tasks that require temporal dependencies, such as sentiment analysis and financial forecasting. In view of the data to be utilised for this study, the logistic regression and random forest classifiers, along with the LSTM network, have been selected for stock price prediction.

3.4 Conclusion

This chapter provided the theoretical and technical foundation necessary for understanding sentiment analysis and machine learning in the context of social media and stock market data. With this foundation in place, the next chapter will focus on establishing the methodology for this study.

CHAPTER 4

Methodology



4.1	Introduction	90
4.2	Data collection.....	91
	4.2.1 Stock selection and historical stock prices	91
	4.2.2 Stock media data	92
4.3	Data pre-processing.....	95
	4.3.1 Feature extraction.....	95
	4.3.2 Label assignment	96
4.4	Model selection	97
4.5	Model validation	100
	4.5.1 Sentiment-driven machine learning model	100
	4.5.2 Analytical methods for relationship analysis.....	104
4.6	Formulation and formalisation of the trading strategy	107
4.7	Model and trading strategy testing	109
4.8	Conclusion.....	111

4.1 Introduction

The methodology for this study is based on a modified version of the CRISP-DM process model, adapted to support a structured development lifecycle for trading systems. The key stages in modelling stock movement based on investor sentiment include data collection, feature extraction, label assignment, model selection, training, evaluation, and deployment (Hernandez-Suarez *et al.*, 2018; Hidayati *et al.*, 2022). Each of these stages is described in this chapter,

expanding on how insights from literature and data availability influenced the methodology for this study.

4.2 Data collection

The data required for this study is twofold. Historical stock price data is needed for evaluating the trading strategy, and social media data relating to news in South Africa is required for sentiment analysis.

4.2.1 Stock selection and historical stock prices

Machine learning models generally perform better on stocks exhibiting specific characteristics related to liquidity and volatility (Cakici & Zaremba, 2024). Investor sentiment, both local and global, has a significant positive correlation with stock market liquidity (Dash *et al.*, 2020; Lee, 2024). Additionally, volatile stocks are more affected by investor sentiment, which also leads to increased co-movement among stocks (Chue *et al.*, 2019; Duxbury *et al.*, 2022). Liquidity refers to the ease with which a stock can be bought or sold without affecting its price, which can be observed in the volumes of the stock traded (Wang, 2024). Volatility represents the degree of variation in a stock's price over time (Gupta, 2024). Stocks with these attributes tend to provide rich and dynamic data patterns, making them well-suited for predictive modelling. Furthermore, stock price forecasting that is based on machine learning techniques performs better in the financial sector, particularly with Long Short-Term Memory (LSTM) models (Burri & Mukku, 2024; Hansun & Young, 2021).

The criteria used for selection are based on data availability from January 2018 to June 2024, their inclusion in the financial sector, and stock-specific volatility and liquidity. Since data availability and sector classification have already been accounted for, the remaining criteria to be assessed are volatility and liquidity. A multi-criteria decision-making method, called linear aggregation, was used to evaluate these two metrics, whereby each stock was ranked individually based on its relative volatility on closing prices and liquidity levels. These ranks were then summed up to generate a composite score for each stock that allows for a balanced consideration of both factors. Accordingly, stocks with the lowest composite scores, reflecting greater volatility and higher liquidity, were selected for subsequent analysis.

The *Yahoo Finance Hub* is an online platform where users can access financial news, data and a variety of analysis tools (Tarale, 2025). Historical data for companies listed on the Johannesburg Stock Exchange was retrieved using the application programming interface (API) of Yahoo Finance (Kumar *et al.*, 2024). Key features available to the public through this website include stock and market data, news and analysis, portfolio tracking, interactive tools, and investment

education. Historical data relating to specific stock price movements corresponding to the time-period of the social media comments was collected from this source. Rather than analysing full sectors or indices, this study focused on a targeted selection of stocks identified by their liquidity and volatility, ensuring their suitability for developing a sentiment- and emotion-driven machine learning trading framework.

4.2.2 Social media data

Social media platforms enable users to create and share content, interact with one another and participate in online communities (Aichner *et al.*, 2021). News outlets are increasingly using social media platforms to strengthen their presence and improve news production (Blank *et al.*, 2012; Hendrickx & Vázquez-Herrero, 2024). Online newspapers are becoming like social media platforms by adopting features such as comments, likes and shares, which enhance user engagement and facilitate the sharing and promotion of news content (Thapaliya, 2022). However, newspapers have numerous advantages over traditional social media posts when it comes to data analysis.

Information could be collected from the internet using methods such as web mining, APIs, Really Simple Syndication (RSS) feeds, downloads from official sources, public repositories, manual collection, and browser automation functions (Bechmann & Lomborg, 2014; Chandra *et al.*, 2019; Lutsenko, 2024; Matta *et al.*, 2020; Patel, 2023). Each of these methods was suited to different types of data and varying levels of availability. The methods considered suitable for the scope of this study included web mining, APIs, downloads, and manual collection.

Web mining is an automatic acquisition method, but not all social media platforms are designed for it (Aguilera *et al.*, 2020; Johnson *et al.*, 2020). Additional ethical and legal issues, like the availability of personal data, lack of consent, data misuse, and terms of service related to web scraping, made this approach undesirable (Boegershausen *et al.*, 2022; Speckmann, 2021).

The *API* method, also known as extraction or harvesting (Matta *et al.*, 2020; Zhao, 2017), is a back-end interface for researchers or third parties enabled through individual platforms (Boegershausen *et al.*, 2022). APIs ensure that access to structured data from social media platforms and other online sources is acquired in a controlled manner (Boegershausen *et al.*, 2022). Fortunately, data availability through APIs from some widely known platforms was high (Abdelfattah *et al.*, 2016; Abu-gellban *et al.*, 2020; Chandra *et al.*, 2019; Constantinescu, 2023; Fottner *et al.*, 2022; Hora & Job, 2024). It was, therefore, crucial to investigate the collection opportunities for some of the most well-known social media platforms.

- The *Facebook Graph API* can extract information related to user profiles, comments and reactions to publicly available posts on Facebook (Abbas *et al.*, 2022). Sentiment analysis is

not widely used on this platform due to various restrictions and security policies concerning data access (Casanez-Ventura *et al.*, 2023). A user account is required (Chandra *et al.*, 2019) and connection requests need approval, increasing bias (Sacramento, 2021). Many spelling mistakes and irregular message lengths necessitate extensive datasets (Drus & Khalid, 2019) which led to time-consuming model training (Biswal *et al.*, 2018). However, smaller datasets could lead to lower classification accuracy (Barz *et al.*, 2022).

- *Google Trends* offers insights into the popularity of top search queries on Google and receives significant attention among researchers in various fields (Cebrián & Domenech, 2023). The built-in API could be used to collect data by searching within a specific time and day range (Abu-gellban *et al.*, 2020), sorted into weekly intervals (Brzeszczyński *et al.*, 2024). Google Trends catered sufficiently to geo-location searches (DeNero *et al.*, 2020) but presented issues that made it unreliable for predictions (Cebrián & Domenech, 2023).
- *Instagram* is a social photography platform where users interact by reading, commenting, and sending posts to other users (Rajasekhar, 2022; Tu *et al.*, 2018). A function that makes this platform unique is its tagging⁷ system which identifies related posts with the hashtag symbol, # (Gulbak & Kirci, 2020). Instagram comments could be extracted using web scraping or APIs, but they were limited to 150 comments (Abdelfattah *et al.*, 2016; Rajasekhar, 2022). As recently as 2021, Instagram was still in a relatively early stage of development compared to Twitter/X, which provided faster delivery of multimedia messages (Ayora *et al.*, 2021; Tu *et al.*, 2018).
- *LinkedIn* is a professional networking platform where people can like, share and comment on posts in addition to pursuing employment opportunities (Katole, 2024; Livisha *et al.*, 2018). The latter distinguishes it from other social media networks (Edwards *et al.*, 2015; Katole, 2024). Although web scraping on LinkedIn is legal (Çetinkaya-Rundel & Dogucu, 2021), automated web scraping attempts are actively blocked (Aradhya & Kumar, 2020). Access to the platform's server was denied without permission and authorisation (Mena-Chalco & Padilha, 2024; Zamora, 2019). Notably, this restriction extended to automated data collection, underscoring that public visibility did not equate to unrestricted access (Sjong, 2021; Zamora, 2019).
- The platform *Reddit* allows users to communicate through posts and comments, also interacting with each other through voting and replying (Fiesler *et al.*, 2021). Information on this platform is categorised into "subreddits", which can range from books, movies, and fitness to food (Li, 2023). Unfortunately, the data was littered with emoticons, unusual spelling and

⁷ Tagging is the act of identifying and linking other accounts, topics, places, or items in a piece of content on social media.

internet slang, which made traditional text analysis challenging (Liu & Yan, 2022). The platform showed potential as a data source (Chie *et al.*, 2023) and was supported by various APIs, including the Reddit API and the Pushshift API, that facilitated data access (Fottner *et al.*, 2022). As the platform had been criticised for its content moderation practices, web scraping raised significant ethical and legal concerns (Carlson & Cousineau, 2020; Fiesler *et al.*, 2024; Johnson *et al.*, 2020).

- The primary purpose of *YouTube* is to share videos online, allowing users to express their opinions and feelings about them (Abbas *et al.*, 2022). Comments on YouTube videos could be retrieved through the YouTube Data API (Chalkias *et al.*, 2023).
- *Twitter/X*, surpasses most platforms due to the short nature of messages and unrestricted following system, making it ideal for sentiment analysis research (Khard & Sonawane, 2016; Sacramento, 2021). Tweets were limited to 280 characters, and data could be extracted via an API or web scraping (Abbas *et al.*, 2022; Chandra *et al.*, 2019). Researchers could apply for academic access to the Twitter/X API (Boegershausen *et al.*, 2022). *Tweepy* and *Twint* are Python libraries used for data extraction from Twitter/X, with *Tweepy* supporting geo-location searches despite only 1% of users sharing their locations (Abbas *et al.*, 2022; Shabrina & Qi, 2023). Twitter/X data allowed filtering by topics such as employment and health, depending on some demographic regions (Mabokela & Schlippe, 2022; Speckmann, 2021). *Tweepy* was expensive and restricted data extraction to the previous seven days, while *Twint* offered web scraping capabilities without API knowledge. However, it had a complicated setup and lacked country-specific capabilities (Abbas *et al.*, 2022; Ağrali & Aydin, 2021). Despite these limitations, most internet-related sentiment analyses are generally performed utilising Twitter/X data (Drus & Khalid, 2019; Yadav, 2023).

It remains unclear whether human beings or computers were responsible for creating online content (Speckmann, 2021). Social media posts might be created by AI-driven robots. This, coupled with issues such as sarcasm, could compromise the quality of data. However, statements and comments made in newspapers were typically not sarcastic and more likely to have been generated by humans (Chandak *et al.*, 2023). Newspaper language also did not often include grammatical errors and slang terms (Mathew & Mathew, 2023). Nevertheless, a vast majority of news agencies integrated their reporting with RSS feeds. The RSS protocol was built upon the Extensible Markup Language and assists users in managing aggregated information from different websites (Bouras *et al.*, 2010). Such delivery systems were considered credible and presented additional data collection possibilities with which the quality of data could be assured (Christophides *et al.*, 2011; Imtiaz *et al.*, 2024).

This discussion reinforced the importance of pre-processing methods, as they would improve the data quality for use in predictive models (Aloysius & Prakash, 2019; DeNero *et al.*, 2020; Parvathi

& Ramachandran, 2019). Feature extraction and label assignment are the two primary methods employed to prepare the dataset for model development. These processes formed an essential part of the data pre-processing stage, ensuring that both the input variables and target outputs were properly structured for sentiment analysis and consequent learning.

4.3 Data pre-processing

Feature extraction transforms raw data into a set of features in preparation for use in machine learning algorithms. Label assignment involves defining the target variable by categorising stock price movements based on predetermined criteria. The following sections cover each of these methods in detail, outlining the procedures used to extract meaningful features from sentiment and market data and to assign appropriate labels representing stock price movements.

4.3.1 Feature extraction

Feature extraction, or vectorisation (Humpherys & Lazrig, 2022) is a foundational step in sentiment analysis modelling because it affects classification accuracy, reduces dimensionality, and converts words to numbers, making them understandable to a computer (Dharma & Saragih, 2022). Vectorisation was required as the chosen classifier was a machine-learning algorithm, and these models operated on numerical input rather than raw or unstructured data (Barnes *et al.*, 2019). Data was mapped to a *Term Frequency Inverse Document Frequency* (TF-IDF) weighted vector space (Garg & Sharma, 2022). As a technique derived from language modelling theory, the TF-IDF approach (Ahmed *et al.*, 2024) also captures the importance of words by assigning higher weights to rare words (Asudani *et al.*, 2023). Indah *et al.* (2021) explain that it utilises the tokens generated through tokenisation to determine a feature's significance by considering the frequency of token occurrence in sentences and their distribution across the dataset. This mathematically translates to

$$TF\text{-}IDF(t, d) = TF(t, d) \cdot IDF(t), \quad (4.1)$$

$$\text{with } TF(t, d) = \frac{f_{t,d}}{N_d}, \quad (4.2)$$

$$\text{and } IDF(t) = \log\left(\frac{N}{n_t}\right), \quad (4.3)$$

where t represents the relevant term, d the document, $TF(t, d)$ the term frequency, $f_{t,d}$ is the frequency of t in d , N_d the total number of terms in d , with $IDF(t)$ the inverse document frequency, N the total number of documents, and n_t the number of documents containing t (Gupta *et al.*, 2023).

After feature extraction, the dataset would be reduced to its most informative variables to enhance model efficiency and improve predictive performance. Various feature extraction methods were

considered, but only a subset of these techniques was used in the final model. This is because excessive or inappropriate feature extraction could lead to a loss of contextual nuance or introduce redundancy, which might reduce model generalisability. To ensure high performance and predictive accuracy in sentiment classification and ultimately stock price forecasting, a supervised learning approach was selected for this study. Consequently, the target outcomes or labels had to be clearly defined, as they guided the machine learning model(s) during training and evaluation.

4.3.2 Label assignment

Label assignment is the process of associating each data instance with a predefined output or category. Unlabelled data is frequently utilised in research and practical applications since labelled data is typically unavailable, inaccessible, or too expensive to acquire (Jayarathna & Sazzed, 2021; Kaji *et al.*, 2014; Talaat, 2023). This means that there is no clear dependent variable, or in this case, sentiment (Boukes *et al.*, 2021). Labelled data is typically processed through a supervised learning framework; unlabelled data is often managed using a combination of unsupervised methodologies and semi-supervised learning (Abdullah *et al.*, 2019; Bag *et al.*, 2023). Labelling is considered a crucial pre-processing step for supervised learning (Nirmala & Sengar, 2020).

TextBlob is one of the simplest open-source APIs provided by the Python library for performing specific (Natural Language Processing) NLP tasks (Gujjar & Kumar, 2021). It is easy to use, install, and implement for developers on all levels, and performs well in simple situations (Jamjuntr *et al.*, 2021). *TextBlob* disregards terms it does not recognise and computes sentiment polarity by assigning predefined scores to known words and expressions, applying weighted averages based on extremes and midpoints to determine the overall sentiment score of a text (Abayomi-Alli *et al.*, 2023). Data requires pre-processing, as elements such as punctuation and stopwords are not considered (Ali *et al.*, 2023).

A key advantage presented by *VADER* is its ability to be used on data directly without the need for pre-training (Ajdari & Hoti, 2023). *VADER* is especially effective for analysing social media content (Ağrali & Aydin, 2021; Hattingh *et al.*, 2023; Bonta *et al.*, 2019). However, it does not understand sarcasm because it does not consider the context of a word (Elwirehardja *et al.*, 2023; Gagiano & Marivate, 2023). Gilbert and Hutto (2014) found that *VADER* surpasses humans in the accuracy of assigning sentiment labels manually. Furthermore, this lexicon can detect and capture the intensity of emotions in data (Al-Shabi, 2020). Fortunately, an extensive set of training data is not required when using *VADER* (Keerthi *et al.*, 2020).

Sentiment labelling in this study was performed using the automated tools, TextBlob and VADER, which converted compound polarity scores into discrete sentiment categories. These tools were well-established, computationally efficient, and suitable for large-scale textual data, which made them appropriate for the scope of this research. Once sentiment labelling was completed, the raw sentiment scores were transformed into structured target variables suitable for supervised machine learning. Effective prediction depended on selecting models that aligned with the nature of the sentiment data and the forecasting objective. The models considered in this study were selected based on their ability to handle text-based features and are discussed in the next section.

4.4 Model selection

Once the data has been labelled, a suitable machine learning algorithm could be selected and applied to model stock price movement. This section presents an examination of the machine learning approaches chosen for this study, with a focus on their applicability to sentiment-based prediction in the financial sector.

Shallow machine learning algorithms that are especially popular for sentiment analysis include decision trees, k-nearest neighbour, Naïve Bayes, support vector machines, logistic regression, and random forests (Aftab *et al.*, 2023; Asudani *et al.*, 2023; Duong & Nguyen-Thi, 2021; Jin *et al.*, 2020). Multi-layered neural networks that process and extract features from data are called deep learning models. These models have been widely applied in areas such as image recognition, NLP, and time-series analysis (Forestier *et al.*, 2024). Some examples of deep learning models commonly used for sentiment analysis include the convolutional neural network (CNN), the recurrent neural network (RNN), and the LSTM network, among others. All the listed approaches were discussed in detail in Chapter 3. Therefore, only the advantages and disadvantages will be discussed in this section to establish the methods to be used for this study.

The decision tree classifier is a form of supervised learning that works by splitting data into branches based on decision rules derived from input features, constructing a tree-like structure. Although decision trees do not require extensive data preparation and pre-processing, they are unstable in the sense that they may change due to pruning. Furthermore, a single tree might excel in prediction while exhibiting signs of overfitting.

The k-nearest neighbour algorithm determines the similarity between existing data and new data, classifying the new data into the class with the highest similarity. It is known for its simplicity and ease of implementation. However, it can construct non-linear decision boundaries, making it less ideal for sentiment analysis.

The Naïve Bayes classifier is a probabilistic classification algorithm based on Bayes' Theorem and relies on conditional probability calculations (Kulkarni *et al.*, 2022). This classifier does not need as much training time and is easy and effective to implement. It can address multi-class feature problems, but the estimates may occasionally be inaccurate. While making the simplistic assumption that data is independent, this method is computationally efficient, performing well for text classification with small training datasets. This means that the presence of one term should not impact the presence of another which can potentially result in suboptimal performance. Complex relationships and correlations between data can be disregarded due to the assumed dependency, ultimately causing a sensitivity to variable selection and affecting prediction accuracy.

Support vector machines are commonly implemented for text categorisation, hand-written symbol recognition, tone recognition, image classification and object detection. They perform well in high-dimensional spaces but are ineffective on large datasets. It also takes a long time to train them on large datasets, making model fine-tuning time-consuming and challenging.

Logistic regression is a form of regression analysis used for predicting the probability that a given input belongs to a particular class. This method is effective and easy to use and interpret, making it a favourable choice when it comes to developing a predictive model. It is the most basic model for classification tasks, and therefore, it serves as an appropriate baseline model for this study.

The random forests approach builds ensembles of decision trees that randomly select subsets of features at each split, which refers to the point where the algorithm chooses a variable and threshold to divide the data into two or more branches. Each tree produces a prediction, and the final outcome is obtained by aggregating the results across all trees. This collective decision-making process reduces variance and generally offers more reliable and accurate classification probabilities. A simple diagram of a decision tree showing a split at a node could be included to visually clarify this process.

The CNN is a deep learning algorithm that excels in complex visual tasks such as image classification, segmentation, and object detection, and it extends to video processing, NLP, and speech recognition. It is relatively easier to train when they have fewer parameters compared to deeper or more complex models. A lower parameter count reduces the risk of overfitting, speeds up-training, and allows CNNs to efficiently combine feature extraction with classification, making them more practical for implementing large-scale networks on limited data or computational resources. A CNN learns spatial feature hierarchies through backpropagation, which means it is not ideal for applications requiring predictions that are dependent on sequential data.

In contrast, an RNN is specifically designed to process sequential data. The connections in RNNs form cycles, so information persists across time steps. This makes RNNs effective for tasks involving sequential data, but they are limited when it comes to handling long-term dependencies, noise, and possible complex relationships. Another type of RNN, the LSTM, was explicitly designed to learn from sequential text data with a feature that allows the model to keep a record of past stock prices. The strength of the LSTM network lies in its ability to selectively remember or forget information, making it ideal for tasks that require temporal dependencies, such as sentiment analysis and financial forecasting.

The traditional machine learning approaches, logistic regression and random forests, were selected for this study, supplemented by the LSTM network for handling sequential text data. These models were chosen based on their popularity, effectiveness and relevance to sentiment classification tasks in the financial sector.

Supervised learning models require labelled data to learn and predict, meaning the data must be divided into training, validation and testing sets (Burzykowski *et al.*, 2023). A training set is used to develop the classifier by learning patterns and relationships within the labelled data. Additionally, the validation set is used to fine-tune the model and select optimal hyperparameters (Burzykowski *et al.*, 2023). Furthermore, the testing set is used to assess how well the trained model performs on previously unseen data (Dobbin & Simon, 2011; Forghani *et al.*, 2020). This split helps to ensure that the model generalises well and is not simply memorising the data. Numerous splitting combinations are found in the literature, like 80/10/10, 70/15/15 and 60/20/20 (Bez *et al.*, 2025) or 80/15/5 (Huynh *et al.*, 2024). The concept is illustrated in Figure 4.1.

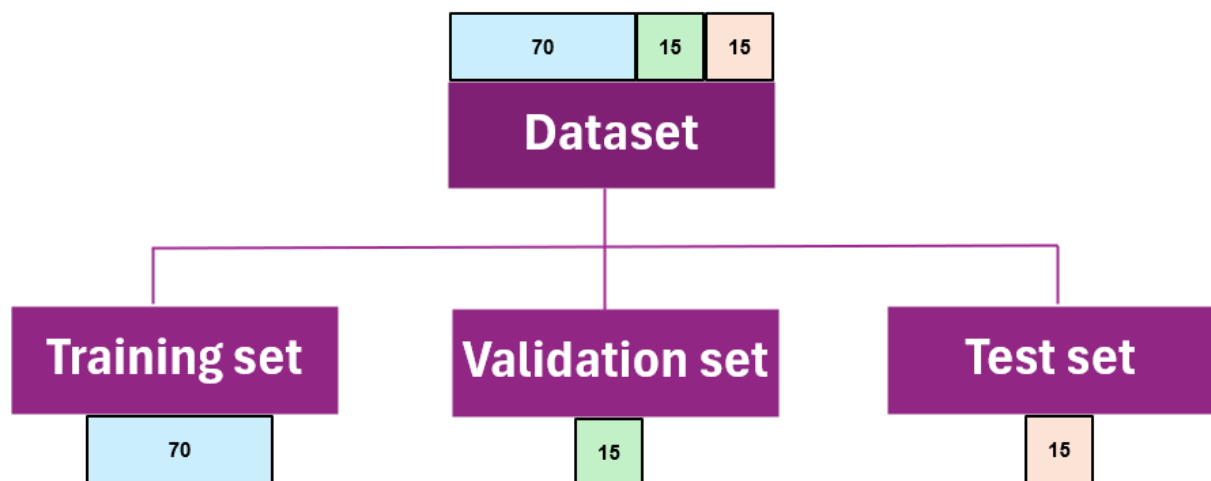


Figure 4.1: Training, validation and testing dataset split

This approach ensures that the model can make reliable predictions for future data, a key aspect in forecasting tasks. The training set will be used to train and develop the machine learning model,

the validation set will be used to validate its performance, and the testing set will be used to evaluate the model as part of the overall trading strategy performance after the trading strategy has been formulated and formalised. Choosing a splitting combination depends on the size and nature of the dataset and can only be done once data collection is complete. Therefore, the splitting of the dataset will be elaborated on after data collection.

4.5 Model validation

The evaluation of results is essential for determining both the statistical robustness and the practical applicability of the proposed framework. This section presents the methods and criteria used to assess model performance and validate outcomes. The evaluation process was conducted in two dimensions. First, the machine learning models were evaluated using established performance metrics to assess their predictive accuracy and stability. Second, analytical methods were employed to investigate the relationship between sentiment indicators and stock market dynamics, ensuring that sentiment-driven outputs held explanatory and predictive value in a financial context.

4.5.1 Sentiment-driven machine learning model

Models can be trained and evaluated multiple times on different combinations of training/validating/testing subsets to ensure reliable model performance (Nawaz, 2023). It may help to identify models that perform well on training data but poorly on test data, a phenomenon known as *overfitting*. In some cases, setting aside large sets of data for testing can reduce the amount of training data, but a process called cross-validation allows each data point to be used for both training, validating, and testing across different splits or folds. A higher number of folds leads to more generalised results (Agrawal *et al.*, 2015). Furthermore, it provides a fair comparison of different machine learning models by evaluating each one multiple times and reducing the impact of random data splits.

Generally, the performance of any classifier is assessed using four key metrics that are summarised in a confusion matrix. Consider the template for a binary confusion matrix in Table 4.1.

Table 4.1: Example of a binary confusion matrix (Doulamis *et al.*, 2021; Kanozia *et al.*, 2023)

		Predicted class		Row total
		Positive	Negative	
Actual class	Positive	TP	FP	$TP + FP = P$
	Negative	FN	TN	$TN + FN = N$
				$P + N = T$

A confusion matrix is a table summarising a classification model's performance based on performance measures (Dhabarde, 2019). Assuming a target variable may be categorised as “Yes” and “No”, Sathyanarayanan and Tantri (2024) define the metrics as follows:

- True positives (TP): the real target value is "Yes" and the model has made a prediction of "Yes".
- False negatives (FN): the real target value is "Yes", whereas the model predicted "No".
- False positives (FP): the real target value is "No", whereas the model predicted "Yes".
- True negatives (TN): the real target value is "No" and the model has made a prediction of "No".

Performance measures derived from a confusion matrix represent the relationship between accurately and inaccurately predicted observations (Agrawal *et al.*, 2015; Akhtar *et al.*, 2019; Sathyanarayanan & Tantri, 2024). These measures include:

- *Specificity*, also known as the true negative rate, measures the proportion of actual negatives that are correctly identified as such by the model. It is a performance metric used to evaluate how well a model distinguishes negative instances (Agudelo *et al.*, 2021).
- *AUC*, or Area Under the Curve, often refers to the area under the Receiver Operating Characteristic (ROC) curve (Dong *et al.*, 2024). The *ROC* curve represents the diagnostic ability of a model, plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The AUC value ranges from 0 to 1, with higher values indicating better model performance. An AUC of 0.5 suggests no discriminative power, equivalent to random guessing, while a value of 1.0 indicates perfect class discrimination. The present concordance measure is the proportion of pairs where the model correctly assigns a higher score to the positive instance than to the negative one. The percentage of ties measures the proportion of pairs where the model assigns the same score to both positive and negative instances. In summary, the estimated AUC provides an overall measure of the model's discriminative ability.

- *Accuracy* is the ratio of correctly predicted observations to the total number of predicted observations (Dong *et al.*, 2024). It becomes misleading if one class significantly outweighs another. The mathematical expression for accuracy is as follows:

$$Accuracy = \frac{TP + TN}{P + N}$$

- The average proportion of correctly predicted positive observations out of the total positive observations is referred to as *precision* (Chicco & Jurman, 2023). Let *TP* denote true positive, *FN* false negative, and *FP* false positive. Precision is useful when false positives are costly and are mathematically formulated as:

$$Precision = \frac{TP}{TP + FN}. \quad (4.5)$$

- Also defined as sensitivity, *recall* quantifies the average proportion of correctly identified positive observations compared to all positive ones (Dong *et al.*, 2024). Determining recall is important when false negatives are risky:

$$\begin{aligned} Recall &= \frac{TP}{TP + FP} \\ &= 1 - \frac{FN}{T}. \end{aligned} \quad (4.6)$$

- The *F1-score* is a fixed weighted average for precision and recall, and it merges the two into a single metric (Dong *et al.*, 2024). While accuracy is more straightforward to grasp conceptually, the F1-score tends to be more beneficial when dealing with imbalanced class distributions. The F1-score balances precision and recall by allowing the trade-off between the two:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (4.7)$$

The three sentiment classes in this study are “Negative”, “Neutral”, and “Positive”. A confusion matrix in such an instance would take the form indicated in Table 4.2.

Each row of the matrix corresponds to the actual class, while each column corresponds to the predicted class. The diagonal elements represent the TPs, where the predicted class matches the actual class. Off-diagonal elements show errors: FPs, where an incorrect class is predicted and FNs, where the true class is missed.

Table 4.2: Expanded multiclass confusion matrix (Afaq *et al.*, 2020; Doulamis *et al.*, 2021)

		Predicted class			
		Class 1	Class 2	...	Class N
Actual class	Class 1	Class (1,1)	FP	...	Class (1,N)
	Class 2	FN	TP	...	FN
	...	...	...	...	...
	Class N	Class (N,1)	FP	...	Class (N,N)

The matrix allowed assessing how well the model distinguished between different classes and identifying where it struggled, providing insights for improving classification accuracy. For instance, by examining the patterns of FP and FN, it could be determined which classes were being confused with others. This detailed analysis was vital for models handling complex datasets with multiple categories.

Similar performance metrics were derived from the multiclass confusion matrix rather than the binary confusion matrix (Doulamis *et al.*, 2021; Doyle *et al.*, 2022; Fahmy, 2023). These measures included:

- *Accuracy* measured the overall performance of the model by dividing the sum of true positives for all classes by the total number of predictions across all classes (Doulamis *et al.*, 2021). It provided a general sense of how often the model correctly classifies samples:

$$Accuracy = \frac{\sum_{i=1}^N TP(C_i)}{\sum_{i=1}^N \sum_{j=1}^N C_{i,j}}, \quad (4.8)$$

where $C_{i,j}$ denoted the entry in the confusion matrix that represented the number of instances whose true class was i and that was predicted as class j . The term C_i refers to the correctly predicted observations for class i , which were the diagonal elements of the confusion matrix ($C_{i,i}$). Accordingly, $TP(C_i)$ indicated the number of true positives for the class i . The variable N represented the total number of classes, while i and j were indices that iterate over the classes from 1 to N .

- The *recall* for a class C_i focusing on the model's ability to correctly identify all instances of that class, was calculated as the proportion of true positives to the sum of true positives and false negatives (Doyle *et al.*, 2022). It is formulated as:

$$Recall \text{ of Class } C_i = TPR(C_i) = \frac{TP(C_i)}{TP(C_i) + FN(C_i)}. \quad (4.9)$$

- *Precision* (or positive predictive value, PPV) for class C_i measured how many of the predicted positives were actually correct, and it was calculated as the ratio of true positives to the sum of true positives and false positives (Doulamis *et al.*, 2021):

$$\text{Precision of Class } C_i = PPV(C_i) = \frac{TP(C_i)}{TP(C_i) + FP(C_i)}. \quad (4.10)$$

- Finally, the *F1-score* for class C_i provided a harmonic mean of precision and recall, balancing both metrics to comprehensively evaluate the model's performance for each class. This is especially useful when dealing with imbalanced datasets (Doyle *et al.*, 2022). The mathematical expression is as follows:

$$F1 \text{ of Class } C_i = F1(C_i) = \frac{2 \times TPR(C_i) \times PPV(C_i)}{TPR(C_i) + PPV(C_i)}. \quad (4.11)$$

The provided equations represented key metrics used for evaluating and validating the performance, stability and generalisation of a multiclass classification model, like those used in this study. However, while the performance metrics provided a clear indication of the predictive capabilities of the machine learning models, they did not fully capture the interaction between sentiment and underlying market dynamics. Analytical methods were employed to address this. They enabled exploration of the relationships between sentiment signals and stock price behaviour.

4.5.2 Analytical methods for relationship analysis

In addition to evaluating model outputs, it was necessary to identify and understand the specific keywords that contributed most strongly to sentiment intensity in the South African context. While these keywords were not directly incorporated into the final trading strategy, their analysis provided valuable interpretive insight into the linguistic and thematic drivers of sentiment formation. This understanding supported a more informed assessment of how market-related discourse reflected underlying investor attitudes and how such sentiment trends may have preceded or coincided with stock market movements. Consequently, the keyword analysis served as a critical diagnostic component that enhanced the interpretability and contextual grounding of the sentiment-based modelling framework.

The extracted keywords were mapped onto established business analysis frameworks, namely political, economic, social, technological, environmental, and legal (PESTEL) (Yüksel, 2012) and Porter's Five Forces (Porter, 2008). The PESTEL analysis categorised keywords according to external macro-environmental factors, whereas Porter's Five Forces framework grouped keywords based on competitive market forces, including industry rivalry, new entrants, supplier power, buyer power, and substitute threats.

Additionally, it was essential to acknowledge the influence of extraordinary global events that have reshaped both investor sentiment and market dynamics (Al-Absy *et al.*, 2024). One such event is the COVID-19 pandemic, which has exerted significant economic, social, and

psychological effects on the South African market environment (Ilesanmi & Olaniran, 2021; Ismail *et al.*, 2022). Considering its far-reaching impact, it was therefore crucial to examine how COVID-19 affected the relationship between sentiment formation and stock price movements. While the pandemic is not explored in detail here, its implications are revisited in Chapter 7 to contextualise the observed sentiment–market interactions during periods of heightened uncertainty.

Analytical methods were therefore employed to assess the extent and nature of the relationships between sentiment and market behaviour. Three complementary techniques were utilised: correlation analysis, lag analysis, and price movement analysis. Collectively, these methods provided important insights into the predictive and explanatory power of sentiment within the South African stock market.

4.5.2.1 Correlation analysis

Correlation analysis examines the statistical relationship between market sentiment and various metrics of stock price movement. The goal was to determine whether positive, negative or neutral sentiment correlated with specific stock price behaviours. The Pearson correlation coefficient and the Spearman rank correlation coefficient are the two commonly used methods for calculating this statistical relationship (Alnafea & Chebbi, 2022; Biehl *et al.*, 2023; Bolek & Bolek, 2024). However, the Pearson correlation coefficient requires normally distributed samples, whereas Spearman's rank correlation is a strong alternative when this assumption is not met (Atukalp, 2021). Moreover, the Spearman Rank correlation coefficient is predominantly used in literature for analysing the relationship between stock price movements and market sentiment (Alamsyah *et al.*, 2021; Chang & Zhang, 2023; Setiawan & Sina, 2023). The coefficient analyses the relationship between two variables by considering the rank of each variable rather than its actual value (Polak, 2021). It is generally carried out by ranking the value of the X variable and the Y variable which, in this study, represented the sentiment and the stock price movement, respectively (Alamsyah *et al.*, 2021). The mathematical representation is expressed in Equation 4.12 (Al-Hameed, 2022)

$$\rho = 1 - \frac{6 \sum_{i=1}^N (R_{X_i} - R_{Y_i})^2}{N(N^2 - 1)}. \quad (4.12)$$

The correlation coefficient, ρ , measures the strength and direction of the relationship between two ranked variables. In this formula, R_{X_i} and R_{Y_i} represent the ranks of the variables X and Y , respectively. The number of observations in the dataset is denoted by N . The coefficient ranges from -1 to 1 . These limits indicate a strong negative or positive correlation, respectively, while values near 0 suggest little to no correlation. This non-parametric method is particularly useful for analysing relationships between variables without assuming normality or linearity in the data. Spearman's Rank correlation coefficient can be interpreted as in Table 4.3.

Table 4.3: Correlation interpretation summary (Setiawan & Sina, 2023)

Correlations	
Correlation value	Correlation interpretation
$\rho = 0$	No correlation
$0 < \rho \leq 0.19$	Very weak correlation
$0.20 \leq \rho \leq 0.39$	Weak correlation
$0.40 \leq \rho \leq 0.59$	Moderate correlation
$0.60 \leq \rho \leq 0.79$	Strong correlation
$0.80 \leq \rho \leq 0.99$	Very strong correlation
$ \rho = 1$	Monotonic correlation

A coefficient of $|\rho| = 1$ signifies a perfect monotonic relationship where the ranks of one variable are a consistent function of the ranks of the other. This classification helps quantify the strength and significance of relationships in statistical analyses.

4.5.2.2 Lag analysis

Lag analysis is a method used to examine how changes in sentiment may affect stock prices over time, but not immediately. Sometimes, it takes time for shifts in public opinion or news stories to impact the market. In lag analysis, researchers aim to determine the time gap, or "lag", between when sentiment changes and when stock prices respond to those changes. For instance, the sentiment of investors might improve in one month due to positive news or trends. However, the impact of this shift on stock prices may not be immediately apparent. Instead, the stock prices may begin to change a few months later. Lag analysis helped to measure this delay and understand how far in advance sentiment could serve as a signal for predicting future stock price movements. By identifying this lag, more informed decisions could be made about when to act on sentiment signals in the market.

Monthly aggregation was selected to stabilise the sentiment index and mitigate sparsity arising from uneven daily headline distributions. While financial markets may respond immediately to breaking news, the objective of this study was to analyse structured medium-term sentiment absorption and persistence effects rather than high-frequency trading reactions. The lag analysis therefore captures the duration and diffusion of sentiment influence rather than the initial reaction timing.

4.5.2.3 Price movement analysis

Price movement analysis focuses on the movement of stock prices and their percentage changes in relation to market sentiment. This includes the following:

- Opening price movement - monthly net changes in the opening prices.
- Closing price movement - monthly net changes in the closing prices.
- Opening price movement percentage - percentage in monthly changes in the opening prices.
- Closing price movement percentage - percentage of monthly changes in the closing prices.

The price movement calculations are discussed in more depth in Section 5.3.1, where the transformation of raw stock prices into monthly movement and percentage change variables is explained. These calculations form a critical part of the analysis, as they establish the foundation for linking sentiment signals to subsequent market behaviour.

The application of correlation, lag, and price movement analyses offers valuable information on how sentiment connects to financial market patterns, complementing the evaluation of machine learning models. Therefore, the next step involves deploying the model in a production environment, where additional considerations play a critical role in ensuring sustained performance and reliability.

4.6 Formulation and formalisation of the trading strategy

The formulation of the dynamic trading strategy integrated the findings and patterns derived from emotion-based sentiment analysis into a structured, data-driven decision-making process. Identifying the emotions underlying each sentiment category formed the foundation for this approach, as emotional intensity and variability provided early indicators of potential market behaviour shifts. Emotion analysis was conducted using the *NRCLex* lexicon, which categorised textual expressions into eight primary emotional dimensions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. These emotional scores were aggregated into a structured dataset for quantitative examination of emotional tone across sentiment categories. Descriptive statistics that include minimum, maximum, mean, standard deviation, and variance were calculated for each emotion within each sentiment class to determine the intensity and stability of emotional expression. This analysis revealed the emotional composition of different sentiment categories. It established the groundwork for clustering analysis, which served to uncover patterns of emotional co-occurrence that may drive collective market sentiment.

The k-means clustering algorithm was employed to partition the emotion-sentiment data into homogeneous groups that share similar behavioural characteristics. This technique is the most popular clustering method that aims to maximise the expected similarity between data points and their corresponding cluster centroids (Enogwe *et al.*, 2021). It is widely used for unsupervised classification, designed to partition data into k distinct clusters based on similar behaviour (Ahmed *et al.*, 2020). However, it is essential to standardise the data into the same ranges (Mohamad & Usman, 2013). The algorithm iteratively minimises the within-cluster sum of squared Euclidean

distances (SSE) between data points and their corresponding cluster centroids, thereby ensuring compact and well-separated clusters (Gill & Singh, 2013). So, k centroids are randomly initialised, data points are assigned to the nearest centroid, and centroids are recalculated as the mean of their assigned points. The process repeats until cluster assignments stabilise or convergence is achieved (Khotimah *et al.*, 2018). Figure 4.2 illustrates the concept of k-means clustering.

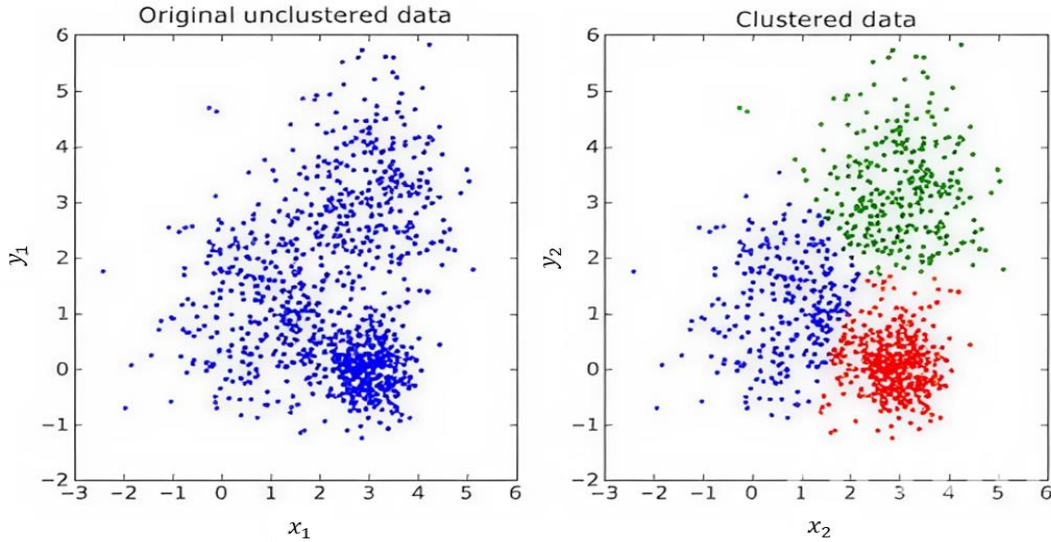


Figure 4.2: Visualisation of the k-means clustering algorithm (Abbas *et al.*, 2018)

The left panel illustrates the dataset before clustering, where all data points are plotted along the x_1 and y_1 axes and uniformly coloured, indicating the absence of identifiable groupings or structure. The right panel displays the outcome after applying the k-means algorithm with an optimal number of clusters ($k = 3$). Here, data points are segmented into three distinct groups, each represented by a different colour (red, blue, and green) along the x_2 and y_2 axes. This visualisation highlights how k-means effectively uncovers hidden structures and natural groupings within the data, forming the foundation for identifying statistically distinct emotion clusters used in subsequent trading strategy formulation.

The elbow method was used to visually determine the value of k as the optimal number of clusters (Gunawan *et al.*, 2020). This technique involved plotting the SSE for various values of k , with the optimal value identified at the "elbow" point of the curve, where the marginal improvement in SSE diminishes (Nainggolan *et al.*, 2019). An example is presented in Figure 4.3.

The x -axis denotes the number of clusters, while the y -axis shows the total within-cluster sum of squared distances (often referred to as "total distance" or "inertia"). As the number of clusters increases, the total distance decreases because data points are assigned to clusters that are closer to them. However, after a certain point, the rate of improvement sharply diminishes. This inflection point represents the threshold where further increases in the number of clusters produce

negligible reductions in total distance. Hence, four clusters were selected as the optimal number for this example, balancing model simplicity with clustering accuracy. This method ensured that the resulting clusters were both meaningful and distinct, avoiding overfitting or unnecessary fragmentation of the data.

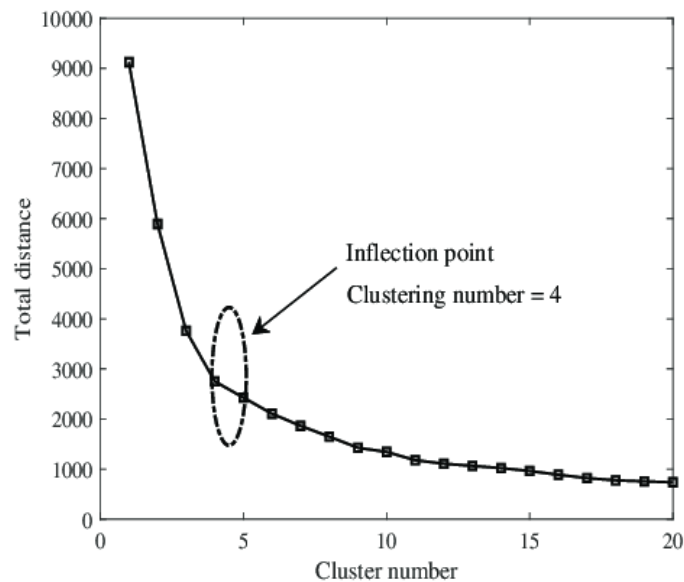


Figure 4.3: Example of the elbow method (Fan *et al.*, 2016)

Following the clustering process, emotion-based thresholds were defined to distinguish normal market sentiment variation from statistically significant emotional deviations. The thresholds were established using a standard deviation margining approach, where the upper and lower boundary limits were derived from the mean and standard deviation within each emotion cluster. These boundary values represented the critical inflection points at which emotional intensity surpassed typical variation, signalling abnormal shifts in market sentiment.

The resulting thresholds served as input parameters for trading signal generation. If a sentiment-emotion value exceeded the upper boundary, it indicated an overexpression of positive emotion and could trigger a potential 'buy' signal. Values falling below the lower boundary suggested a rise in negative emotions, which could activate a 'sell' signal, while intermediate values corresponded to 'hold' decisions. This mechanism translated emotional dynamics into actionable market signals, enabling adaptive responses to evolving sentiment conditions.

4.7 Model and trading strategy testing

The performance of the sentiment-driven trading signals was evaluated through a structured testing framework designed to assess their predictive validity, responsiveness, and reliability in a

simulated market environment. The testing phase focused on evaluating the practical performance of the model-generated trading signals using the test split of the dataset and historical market data corresponding to the selected stocks identified in Section 4.2.1.

Only stocks with available lag-adjusted sentiment predictions were included to ensure temporal consistency between sentiment generation and market reaction. Prior to testing, a simplified linear aggregation method was applied to evaluate liquidity and volatility. Each stock was ranked according to its relative liquidity and closing price volatility, and these ranks were combined to produce a composite score that balanced both factors. This ensured that the stocks selected for testing represented those with sufficient trading activity and price variability to reflect sentiment-driven market behaviour meaningfully. The testing framework replicated real-world trading conditions by applying the sentiment-emotion-driven buy, hold, and sell signals to historical price data over the defined testing period. Each signal was aligned to the appropriate lag structure, representing the delay between sentiment generation and price response, thereby allowing the assessment of short-term and delayed market reactions.

The testing process consisted of several sequential steps:

- Signal application: The calibrated trading signals, derived from the emotion thresholds and sentiment polarity classifications, were systematically applied to the stock price data to simulate trading actions.
- Lag structure evaluation: Trading signals were mapped across multiple lag intervals (Lag 0 to Lag 6) to evaluate how predictive performance evolved when sentiment signals were shifted in time relative to market prices.
- Return measurement: For each stock, the test framework computed both individual and cumulative returns, providing a comprehensive assessment of the strategy's predictive capability and its potential to capture directional market trends.

To ensure consistency and comparability, the emotion data was standardised prior to testing, and the same thresholding logic established during the strategy formulation phase, was maintained. This testing design provided a controlled mechanism for analysing how effectively sentiment-emotion signals could anticipate price movements and adapt to different temporal lags.

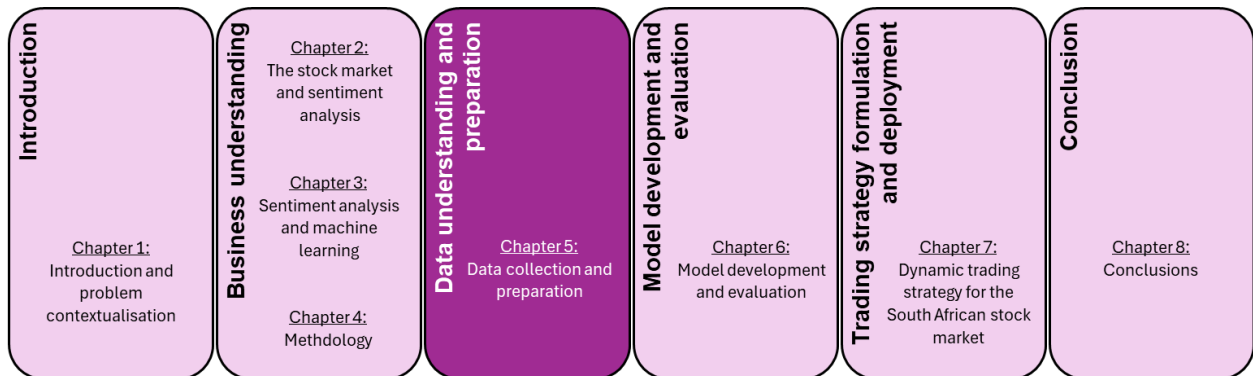
Upon completion of this phase, the sentiment-emotion framework was formalised into a decision-support mechanism capable of generating context-sensitive buy, hold, and sell signals. The subsequent deployment phase focuses on translating this validated model into an operational environment, supporting the real-time execution of sentiment-based trading strategies under dynamic South African market conditions.

4.8 Conclusion

The methodology for developing a sentiment-based trading strategy, guided by a modified CRISP-DM framework, was tailored and discussed in this chapter. Each stage of the process, from data collection to trading strategy formalisation, was designed to align with the study's objectives. The resulting methodology presented a strong foundation for the experimental work that followed. The next chapter will cover the data collection, feature extraction, and label assignment stages.

CHAPTER 5

Data collection and preparation



5.1	Introduction	112
5.2	Data collection.....	113
	5.2.1 Stock selection and historical stock prices	113
	5.2.2 Social media data	125
5.3	Data pre-processing.....	131
	5.3.1 Stock price data.....	131
	5.3.2 Social media data	133
5.4	Conclusion.....	136

5.1 Introduction

The focus of Chapter 5 is on the collection and the preparation of data that was used to train, validate and test the machine learning models for suitable hyperparameter optimisation and evaluation. Data was gathered from two primary sources, namely stock price data for understanding market movements, and social media. The next section covers a rationalisation of the factors considered for the choice of stocks used in this study, followed by the selection of the platform for collecting data from social media and the data collection process. The preparation of stock price data is covered in Section 5.3.1, followed by pre-processing, feature extraction from social media data, and label assignment in Section 5.3.2.

5.2 Data collection

The data collection approach for this study consists of two parts. Historical stock price data was collected from Yahoo Finance Hub (Kavya *et al.*, 2021; Kumar *et al.*, 2024). Accurate sentiment analysis depends strongly on the quality of the data used (Bhalla & Pooja, 2022; Narayan & Samanta, 2023). High-quality data is characterised by accuracy, completeness, relevancy, and representativity (Easton *et al.*, 2024; Hassenstein & Vanella, 2022; O'Donnell *et al.*, 2023). The titles of general news in South Africa were collected from the platforms best suited for this. The collection period spanned from January 2018 to June 2024. However, only the most recent training and validation period was used for descriptive statistics and stock selection to avoid data leakage, and to ensure unbiased supervised learning.

5.2.1 Stock selection and historical stock prices

The study focused exclusively on shares listed on the South African stock exchange, known as the Johannesburg Stock Exchange (JSE). The JSE offers a diverse range of shares across various sectors, including finance, mining, and manufacturing, with further variation in stock prices, trading volumes, and liquidity. Historical data for these companies was obtained on the Yahoo Finance application programming interface (API). Each of the discussed stocks was identified by its unique ticker⁸ symbol, which was used to retrieve stock price data for the specified date range.

The key metrics collected include the opening price, highest price, lowest price, closing price, adjusted closing price, and trading volume, all of which were recorded on the first trading day of each month. The data was extracted, organised into a multi-dimensional arrangement, and transformed into a structured format, suitable for time-series analysis and further processing. An excerpt of the data is shown in Table 5.1.

While the values presented in the table are only for illustration purposes, the same process was applied systematically to each of the selected stocks. While comprehensive market analyses often involve large sector-based portfolios or filtering by price or volume, this study adopted a different approach. Instead of analysing entire sectors or index components, a limited selection of shares was determined based on liquidity and volatility to be appropriate and relevant for constructing a sentiment- and emotion-informed machine learning-based trading framework.

⁸ A ticker is a unique short symbol used to identify a specific publicly-traded stock on an exchange

Table 5.1: Example of historical stock data extracted from Yahoo Finance Hub

Date	Statistics					
	Open (R)	High (R)	Low (R)	Close (R)	Adjusted Close (R)	Volume
31-Jan-23	20000	20775	18588	19845	17544	56284872
28-Feb-23	19547	20423	16734	18178	16070	81630620
31-Mar-23	18218	18884	17150	17764	15704	44983474
30-Apr-23	17764	17797	15070	15383	14104	74720827
31-May-23	15545	17114	15352	16791	15395	80594270
30-Jun-23	16762	19020	16739	18945	17370	61385667
31-Jul-23	18945	19055	17065	18229	16713	55617666
31-Aug-23	18499	18549	16964	17483	16029	47790091
30-Sep-23	17483	18115	16533	16992	16192	36449318
31-Oct-23	17430	18626	16001	17376	16558	56483031
30-Nov-23	17527	17616	15351	16371	15600	62225671
31-Dec-23	16371	16485	15427	16380	15609	51673727

Machine learning models perform best on stocks that are both liquid and volatile, as these characteristics yield richer predictive patterns (Cakici & Zaremba, 2024). Based on data from January 2018 to June 2024, stocks within the financial sector were evaluated using a linear aggregation method that ranked each stock by its volatility and liquidity, producing a composite score to ensure a balanced selection. Stocks with the lowest composite scores, which indicate high volatility and high liquidity, were selected for further analysis. Accordingly, seven stocks were chosen for inclusion in the modelling process: Hammerson PLC, Redefine Properties, FirstRand Limited, Equites Property Fund, Finbond Group Limited, Growthpoint Properties, and Standard Bank Group. The study limits its scope to a manageable group of highly liquid and actively traded stocks to ensure methodological feasibility and reliable signal calibration. Restricting the analysis to a small, well-selected sample ensures computational tractability while preserving high data quality. This approach allows for a clearer interpretation of model performance than would be possible in a large. A summary of this analysis is shown in Table 5.2.

Table 5.2: Summary of the liquidity and volatility analysis

Stock	Volatility and liquidity analysis					
	Liquidity	Volatility	Liquidity rank	Volatility rank	Total score	Final rank
Hammerson PLC	56504836	0.01378	5	4	9	1
Redefine Properties	321736404	0.00791	12	9	21	2
FirstRand Limited	277774843	0.00784	11	14	25	3
Equites Property Fund	39410110	0.01211	1	29	30	4
Finbond Group Limited	314095181	0.00568	3	30	33	5
Growthpoint Properties	66785046	0.00626	16	19	35	6
Standard Bank Group	66785046	0.07910	40	3	43	7

Hammerson PLC achieved the top position with a balanced combination of relatively high volatility and strong liquidity. Redefine Properties and FirstRand Limited followed, both showing favourable liquidity levels but lower relative volatility. This ranking approach allowed for a systematic and transparent selection of stocks for further analysis based on the key criteria. The seven selected stocks are briefly described and motivated next, based on their suitability for inclusion in the modelling framework (Yahoo Finance, 2025).

5.2.1.1 Hammerson PLC

Hammerson PLC was a prominent property development and investment company with a diversified portfolio of retail properties across Europe, as well as South Africa. As a dual-listed entity on the JSE and the London Stock Exchange, it provided cross-border exposure and reflected investor sentiment in the retail and property sectors. The company’s consistent availability of price and volume data supported its suitability for inclusion in sentiment analysis and emotion-driven trading strategies using machine learning approaches. Figure 5.1 illustrates the monthly stock performance and trading volume trends of Hammerson PLC for the year 2023.

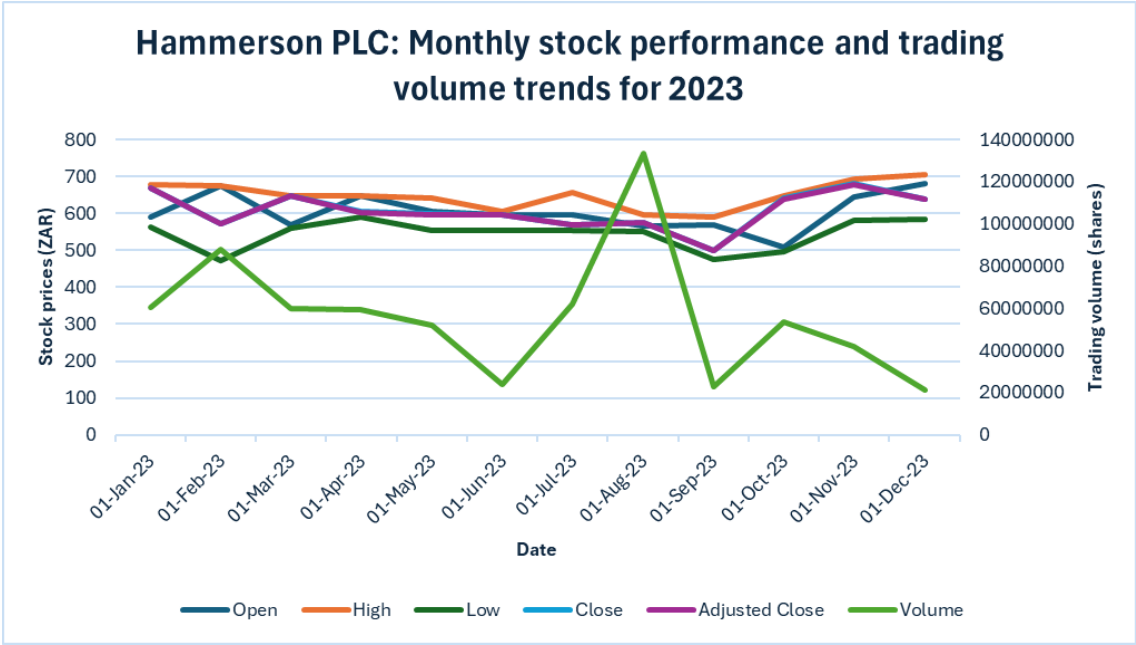


Figure 5.1: 2023 Stock performance and trading volume trends for Hammerson PLC

The stock price metrics remained relatively steady throughout the year, with modest fluctuations. Notably, the adjusted close price closely resembled the closing price, indicating minimal adjustments for dividends or splits. In contrast, the volumes of the stock traded showed considerable variability, with a sharp spike in August, followed by a steep decline in September and a continued downward trend through December. This divergence between price stability and

volume volatility might suggest periods of increased investor activity or a response to specific market events during the mid-year. These trends are summarised in Table 5.3.

Table 5.3: 2023 Statistics on Hammerson PLC

Period	Statistics				
	Opening price variance	Average opening price (R)	Closing price variance	Average closing price (R)	Average volume
3 months	0.02937	612	0.02454	653	38807253
6 months	0.01388	595	0.01824	600	55719059
12 months	0.01384	604	0.01378	608	56504836

The values presented in Table 5.3 provide a summary of Hammerson PLC's stock performance, volatility, and liquidity across three time periods: 3 months, 6 months, and 12 months. Volatility was assessed by calculating the variance of opening and closing prices, while the average trading volume indicated liquidity. The 3-month period exhibited the highest variance in both opening and closing prices, signalling greater short-term price swings and potential market instability. Conversely, the 12-month period showed the lowest variance, implying more stable price behaviour over time. The average trading volume increased from the 3-month to the 12-month period, reflecting enhanced liquidity and sustained investor interest over the longer term.

5.2.1.2 Redefine Properties

Redefine Properties was a well-established real estate investment trust specialising in commercial, retail, and industrial property investments. The company held a significant market presence on the JSE and offered extensive historical data relevant to property valuation, investor activity, and sector-specific trends. As real estate markets often reflect broader economic sentiment, particularly through interest rate changes and urban development trends, this trust was a strong candidate for sentiment-based modelling in the property investment domain. Its sector-specific sensitivity to news, macro-economic signals, and investor sentiment supported its inclusion in this study. The 2023 monthly trends in stock performance and trading volume for Redefine Properties are depicted in Figure 5.2.

Price metrics showed moderate fluctuations throughout the year, with a relatively stable upward trend towards the year-end. The adjusted closing price closely followed the actual closing price, indicating that minimal post-trade adjustments were made. Trading volume saw notable spikes in April and October, possibly reflecting periods of increased investor activity or reaction to company- or market-specific events.

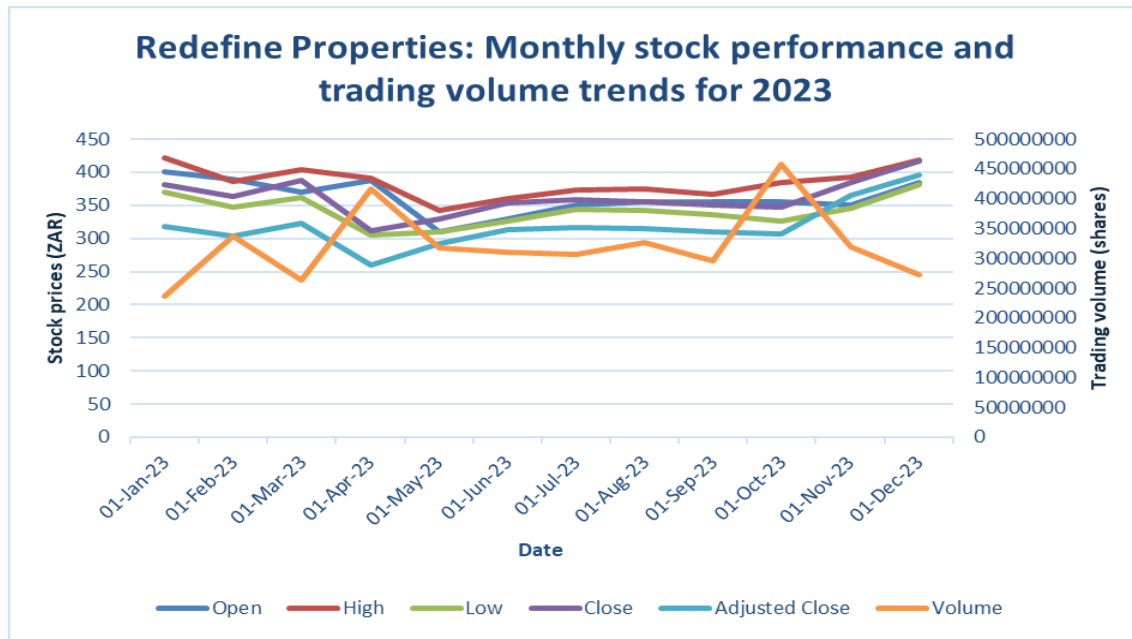


Figure 5.2: 2023 Stock performance and trading volume trends for Redefine Properties

Although there was some fluctuation, the overall volume remained consistently high, supporting the stock’s liquidity and its potential suitability for sentiment-driven trading strategies. Table 5.4 summarises the main data trends.

Table 5.4: 2023 Statistics on Redefine Properties

Period	Statistics				
	Opening price variance	Average opening price (R)	Closing price variance	Average closing price (R)	Average volume
3 months	0.00355	363	0.00351	383	350078928
6 months	0.00182	359	0.00264	369	329944099
12 months	0.00727	362	0.00791	362	321736404

A summary for Redefine Properties across 3-, 6-, and 12-month periods is presented in Table 5.4, highlighting the stock’s strong price stability and high liquidity. The standard deviation of closing prices remained exceptionally low across all timeframes, ranging from R25 to R29 (in South African rand), indicating minimal price volatility. Average closing prices also remained consistent around R362 to R383. This reflected the valuation stability characteristic of real estate investment trusts. In contrast to the low volatility, trading volume was exceptionally high, with an average of over 321 million shares traded monthly over the past year. This high level of liquidity ensured market responsiveness despite the stock’s overall pricing stability. The traits of low volatility, consistent pricing, and significant trading activity positioned Redefine Properties as a prime candidate for sentiment-based modelling in a stable asset class such as property investment.

5.2.1.3 FirstRand Limited

FirstRand Limited was one of South Africa's largest financial institutions, encompassing retail, commercial, corporate, and investment banking services. Its significant role within the local economy and strong trading presence on the JSE made it a prime proxy for financial market sentiment. The company's comprehensive historical datasets and market activity rendered it highly appropriate for machine learning-based sentiment analysis and predictive modelling. Monthly fluctuations in FirstRand Limited's stock performance and trading volume during 2023, are shown in Figure 5.3.

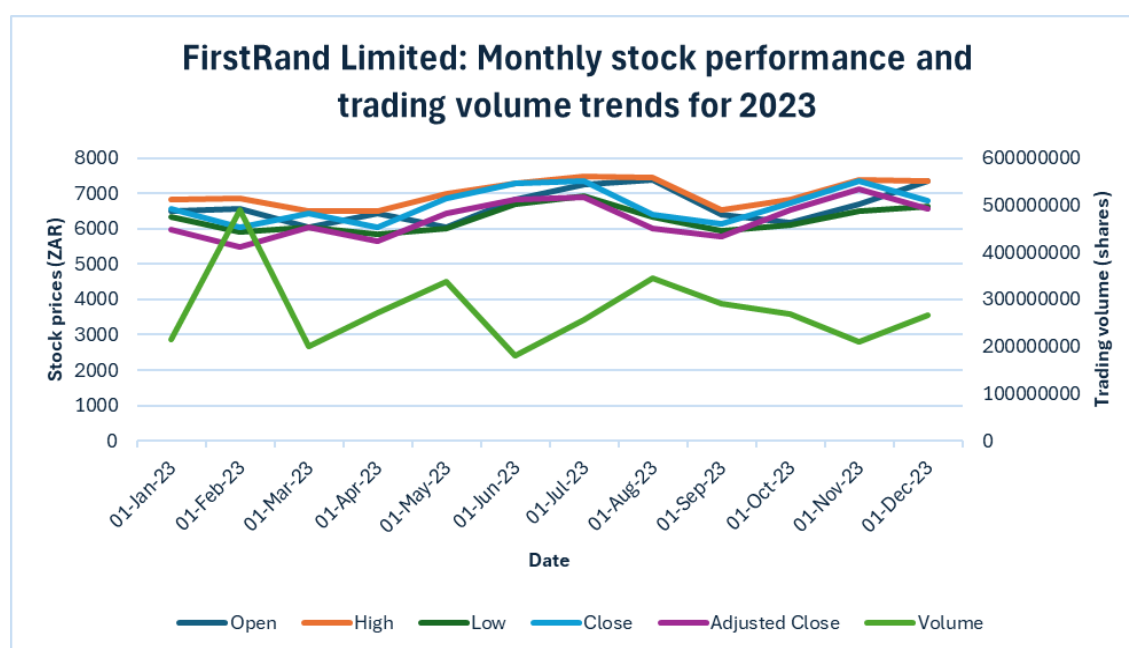


Figure 5.3: 2023 Stock performance and trading volume trends for FirstRand Limited

The stock's price indicators showed relatively stable movement throughout the year, with a slight dip in August and a gradual recovery towards the end of the year. The adjusted closing price tended to align closely with the actual closing price, indicating minimal post-market adjustments. Trading volume showed more noticeable variation, with a sharp peak in February followed by a general downward trend and fluctuations in subsequent months. Despite this variability, the overall volume remained significant, supporting the stock's liquidity and its importance for sentiment-driven trading strategies.

Table 5.5 summarises the volatility, average price levels, and liquidity of FirstRand Limited over three time periods: 3-, 6-, and 12-months. The highest closing price variance occurred in the most recent 3-month period, suggesting increased short-term market fluctuation. However, the 12-month data showed lower and more stable variance values, indicating consistent long-term price behaviour.

Table 5.5: 2023 Statistics on FirstRand Limited

Period	Statistics				
	Opening price variance	Average opening price (R)	Closing price variance	Average closing price (R)	Average volume
3 months	0.00506	6745	0.00951	6962	248242856
6 months	0.00803	6878	0.00850	6793	272680098
12 months	0.00704	6639	0.00784	6662	277774843

Liquidity, as measured by average trading volume, increased across the timeframes, with the 12-month average exceeding 277 million shares per month. This highlighted strong and sustained investor interest over the longer term.

5.2.1.4 Equites Property Fund

The Equites Property Fund was a specialised real estate investment trust focused on logistics and industrial properties, in both South Africa and the United Kingdom. Its sector-specific exposure provided a unique lens into investor sentiment within the logistics property segment. As an actively traded entity with accessible historical trading data, the fund supported the methodological requirements for sentiment and emotion analysis using machine learning. Figure 5.4 presents the stock performance and trading volume of Equites Property Fund every month for 2023.

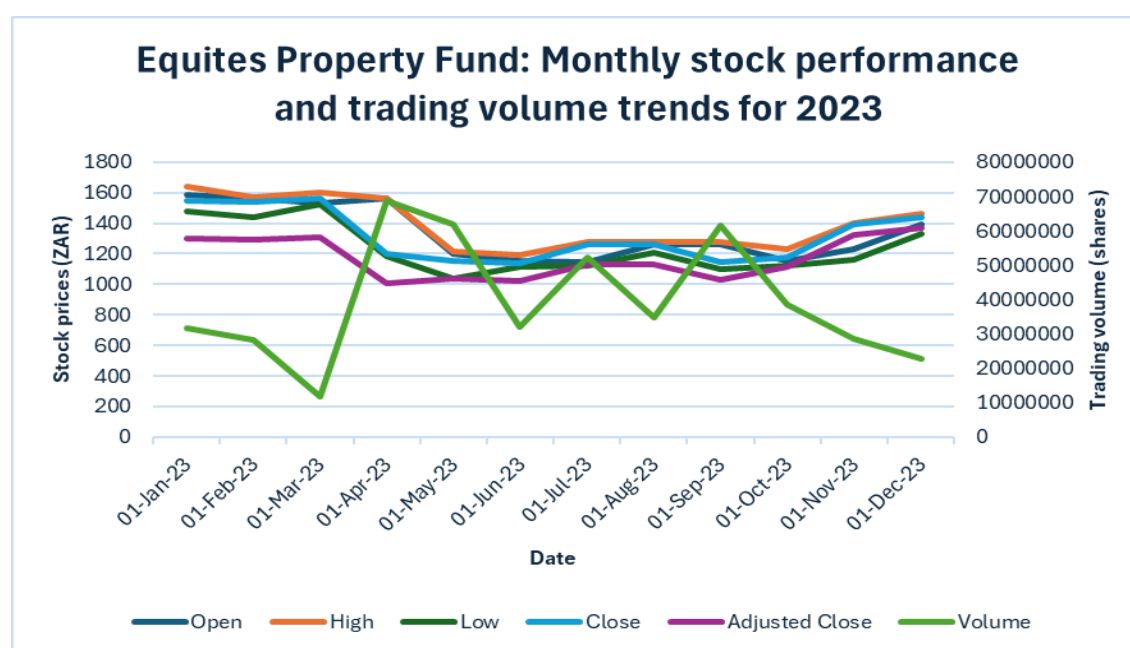


Figure 5.4: 2023 Stock performance and trading volume trends for the Equites Property Fund

From January to June, the stock's open, high, low, close, and adjusted close prices showed a declining trend, followed by a gradual recovery towards December. Throughout the year, the

adjusted close prices remained closely aligned with the closing prices, suggesting that minimal corporate actions had an impact on valuation. The trading volume indicated significant fluctuations, with notable peaks in April and September, and a steep decline in the final quarter. Despite moderate price volatility, the volume trend indicated periods of irregular investor activity, which may affect liquidity and the consistency of execution in sentiment-driven trading strategies. Consider the overview in Table 5.7.

Table 5.6: 2023 Statistics on the Equites Property Fund

Period	Statistics				
	Opening price variance	Average opening price (R)	Closing price variance	Average closing price (R)	Average volume
3 months	0.01226	1259	0.00684	1336	29956127
6 months	0.00626	1240	0.00827	1279	39771247
12 months	0.01060	1337	0.01211	1317	39410110

The highest volatility in the closing price of 0.01211 was observed over the 12-month period. This indicated that there would be larger price fluctuations in the long term. In contrast, the 3-month and 6-month periods showed more stable behaviour in either the opening or closing prices. The average trading volume remained relatively consistent across the 6- and 12-month periods, at approximately 39 million shares, indicating a steady level of investor interest. However, the 3-month average volume was lower, reflecting a possible recent decline in trading activity.

5.2.1.5 Finbond Group Limited

The Finbond Group Limited operated within the alternative finance and micro-lending sector in South Africa and North America. Its focus on short-term lending and niche financial services made it particularly sensitive to shifts in public sentiment, regulation, and economic outlook. Despite being a smaller-capital stock, its trading activity and availability of monthly market data supported its inclusion in sentiment-based predictive models. Figure 5.5 illustrates the monthly stock performance and trading volume trends of the Finbond Group Limited for the year 2023.

Finbond’s open, high, low, close, and adjusted close prices remained fairly stable throughout 2023, with minor fluctuations. The adjusted closing price remained closely aligned with the actual closing price, indicating that adjustments from dividends or corporate actions were negligible. In contrast, trading volume showed a gradual decline, especially in the second half of the year, dropping from its peaks in the early months to noticeably lower levels by December.

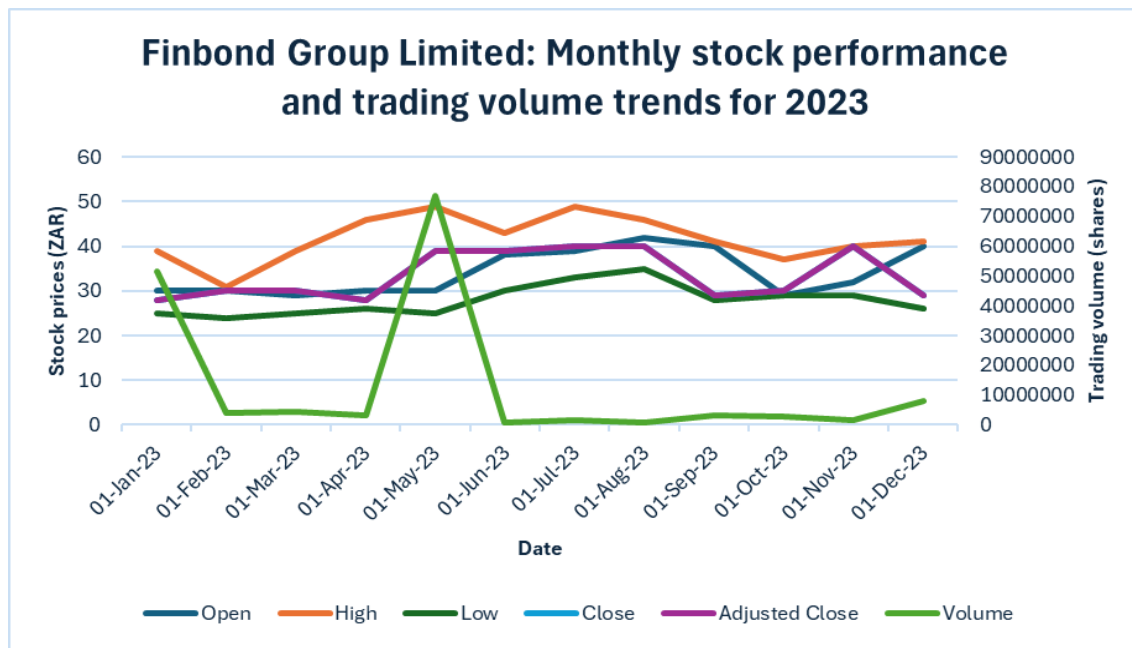


Figure 5.5: 2023 Stock performance and trading volume trends for the Finbond Group Limited

This decline in volume might have been a result of waning investor interest or reduced trading activity. This, in turn, could have impacted the stock's liquidity and the reliability of executing sentiment-driven trades. Key trends are represented in Table 5.7.

Table 5.7: 2023 Statistics on the Finbond Group Limited

Period	Statistics				
	Opening price variance	Average opening price (R)	Closing price variance	Average closing price (R)	Average volume
3 months	0.08145	34	0.09366	33	4175046
6 months	0.03420	37	0.05535	35	3026735
12 months	0.02221	34	0.04106	34	13276229

Volatility was at its highest over the last 3-month period, indicating an increased short-term price instability. In contrast, the 12-month variances meant a more stable price pattern, suggesting a reduced long-term fluctuation. The average opening and closing prices remained relatively consistent across all periods. However, the average trading volume decreased sharply in the shorter timeframes, with only 4.2 million shares traded monthly over 3 months, compared to approximately 13.3 million over 12 months. This trend pointed to a decline in recent trading activity, which may have affected the stock's liquidity and execution reliability in sentiment-based strategies.

5.2.1.6 Growthpoint Properties

One of South Africa's largest real estate investment trusts, Growthpoint Properties, held a diversified portfolio of commercial, retail, and industrial properties. With a substantial market capitalisation and consistent presence on the JSE, the company provided a strong foundation for real estate market representation in predictive modelling. The relatively stable nature of the property sector, combined with the availability of detailed financial disclosures and investment data, enhanced Growthpoint's suitability for machine learning-based sentiment analysis. The pricing and volume patterns were influenced not only by general economic sentiment but also by interest rate decisions, property demand cycles, and market liquidity - all of which could be captured through sentiment indicators derived from textual data. The trends in Growthpoint Properties' stock performance and trading volume over the course of 2023, are displayed in Figure 5.6.

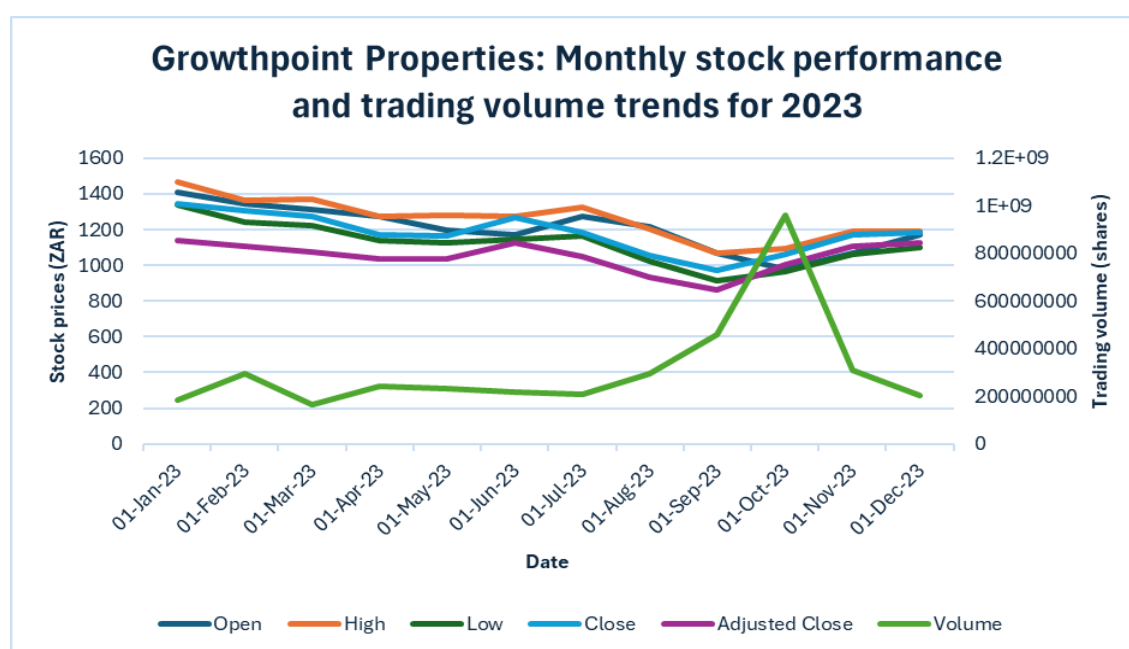


Figure 5.6: 2023 Stock performance and trading volume trends for the Growthpoint Properties

The price indicators showed a mild downward trend in the first three-quarters of the year, with a slight recovery from October. Notably, the adjusted closing price consistently stayed closely aligned with the standard closing price. This suggested consistent data, unaffected by major corporate actions. This was particularly important in property-related stocks, where dividend events could often disrupt valuation benchmarks. The data on trading volume revealed a sharp spike in October, far surpassing its earlier levels, which was followed by a decline in November and December. This surge in investor activity might have been linked to earnings releases or

macro-economic updates related to interest rate adjustments or sector outlook. The observed patterns are summarised in Table 5.8.

Table 5.8: 2023 Statistics on the Growthpoint Properties

Period	Statistics				
	Opening price variance	Average opening price (R)	Closing price variance	Average closing price (R)	Average volume
3 months	0.01083	1072	0.00209	1138	491088593
6 months	0.00989	1129	0.00837	1103	406639973
12 months	0.00540	1207	0.00568	1179	314095181

Volatility is indicated by the variance of the opening and closing prices, and the 3-month period showed a notably higher variance in the opening price of 0.01083. This implied an instability in the short term. In contrast, the variance in both the opening and closing prices decreased over the 12-month period, suggesting a more stable long-term price behaviour. The average opening and closing prices stayed within a limited range, reflecting consistent valuation levels. Liquidity for this stock remained exceptionally strong across all periods, with a peak of over 491 million shares traded monthly during the latest 3 months. Despite some short-term volatility, the stock demonstrated solid and sustained liquidity, supporting its reliability for a sentiment-based trading strategy.

5.2.1.7 Standard Bank Group

The Standard Bank Group is one of Africa's most established and diversified financial institutions, offering services in banking, wealth management, and insurance. Its leadership position and large market capitalisation on the JSE have made it a benchmark for sentiment within the banking and financial services sector. The extensive availability of historical financial and trading data enhances its value as a candidate for sentiment-driven stock prediction models using machine learning. The 2023 monthly trends in stock performance and trading volume for this group are depicted in Figure 5.7.

The stock's price indicators demonstrate a consistent upward trend, accompanied by moderate fluctuations across the year. The adjusted closing price has been tracking the closing price closely, implying that post-trade adjustments are negligible. On the secondary axis, the trading volume is shown, which has varied noticeably over the year, with peaks in February and May. This was followed by a general decline until November and a slight recovery in December. These changes in trading activity might have been linked to significant market events or shifts in investor sentiment.

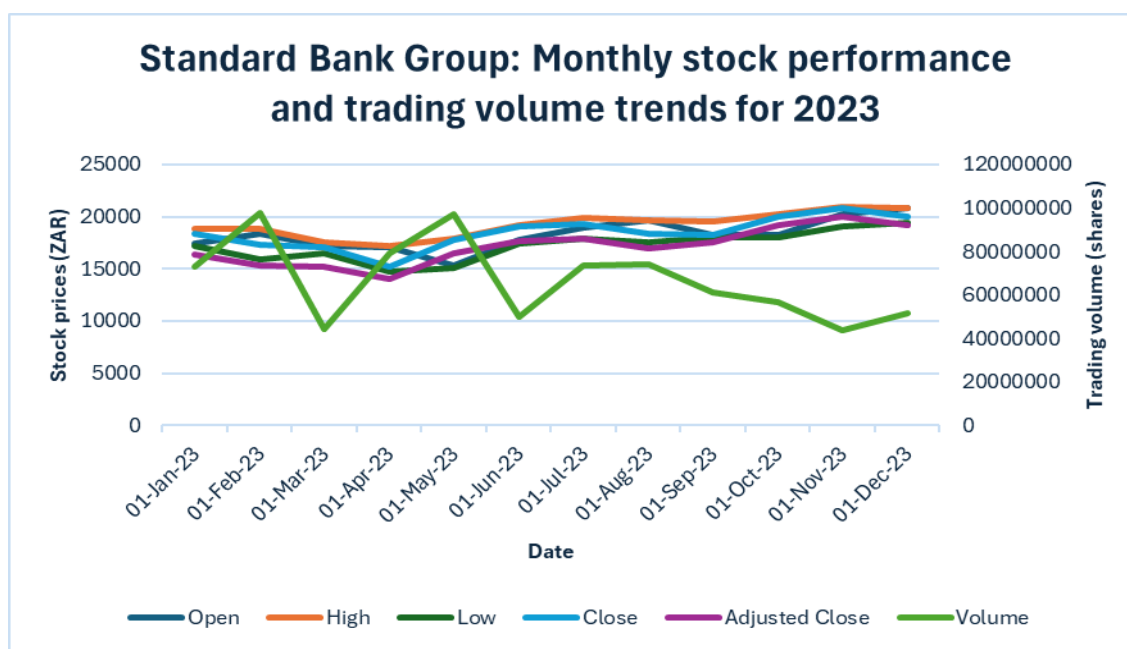


Figure 5.7: 2023 Stock performance and trading volume trends for the Standard Bank Group

Overall, the stock shows both price stability and active market participation, which is relevant for sentiment-based trading assessments. Table 5.9 provides a statistical overview that captures the underlying trends.

Table 5.9: 2023 Statistics on the Standard Bank Group

Period	Statistics				
	Opening price variance	Average opening price (R)	Closing price variance	Average closing price (R)	Average volume
3 months	0.01083	1072	0.00209	1138	491088593
6 months	0.00989	1129	0.00837	1103	406639973
12 months	0.00540	1207	0.00568	1179	314095181

The volatility was at its highest in the short term, with the opening price in the 3-month period recording the most significant variance of 0.01083. However, both opening and closing price variances decreased steadily over more extended periods, suggesting a greater price stability in the medium to long term. The average opening and closing prices showed minor fluctuations, reflecting a consistent stock valuation. This stock's liquidity has remained exceptionally strong, with average monthly trading volumes exceeding 314 million shares over 12 months, peaking at more than 491 million over the 3-month period. These patterns indicate robust investor interest and consistent market activity, which support the stock's suitability for sentiment-based trading strategies.

The stocks discussed in the preceding paragraphs represent the ideal liquidity and volatility combination that has been identified from the lowest composite scores. This has made them suited for use in machine learning to identify predictive patterns and to forecast market trends.

Their inclusion in this study ensures a diverse and representative portfolio that is aligned with the research objectives, providing a reliable foundation for further sentiment integration and predictive analysis.

5.2.2 Social media data

In establishing the methodology for this study, it became evident that the essential factors to consider when selecting a platform for data collection included quality, quantity, availability, consistency, relevance, and legal and ethical concerns. News reported on social media platforms was considered consistent and timely, as they operated continuously, with users active at any time of the day (Benitez *et al.*, 2022). Each platform had constraints in terms of API use as well as legal and ethical considerations. These considerations were compiled into a checklist, which is shown in Table 5.10.

Table 5.10: Checklist for selecting the appropriate data source

Factor	Platform						
	Facebook	Google Trends	Instagram	LinkedIn	Reddit	Twitter/X	YouTube
Data quality	X	X	X	X	X	X	X
Data quantity	✓	✓	✓	✓	✓	✓	✓
Data availability	✓	✓	✓	X	✓	✓	✓
Relevance	X	X	X	X	X	✓	X
Data consistency	✓	✓	✓	✓	✓	✓	✓
Legal and ethical considerations	✓	✓	✓	X	✓	✓	✓

Data quality issues such as slang terms, emoticons and emojis, abbreviations, Uniform Resource Locators (URLs⁹), hashtags, spelling errors, single characters, irregular message lengths, and improper grammar and punctuation are widely recognised as universal challenges in text data (Drus & Khalid, 2019; Musial & Kazienko, 2013; Nandwani & Verma, 2021). While all platforms provide high quantities and consistency of data, only Twitter/X meets the criterion for relevance.

⁹ A URL is a Uniform Resource Locator that represents the address to type into a web browser to visit a specific page.

Availability varies, with LinkedIn lacking open data access. Legal and ethical considerations could also be met for all platforms, except for LinkedIn.

Considering the characteristics of each platform and the fact that extensive pre-processing is expected to improve data quality, data was collected using two complementary approaches: Really Simple Syndication (RSS) feeds and web mining. RSS feeds provided structured Extensible Markup Language (XML) based syndication links that newspapers made available for automated distribution. These feeds typically contained the article title, publication date, and URL, which allowed for systematic extraction across all newspapers that supported this format. This method yielded a dataset of 26,964 valid entries from sources with RSS feeds that were active in August 2024.

For newspapers without comprehensive RSS coverage or where archives were not fully accessible through feeds, web mining was employed. This involved programmatically visiting a newspaper's archive pages, iterating through paginated results, and extracting article titles and dates directly from the HyperText Markup Language (HTML) structure. While technically distinct from RSS feed extraction, this method allowed the inclusion of additional sources and produced a larger dataset of 48,896 entries. Therefore, different algorithms were employed for collecting data from these sources. Algorithm 5.1 was used to extract news information from RSS feeds.

Algorithm 5.1: Data collection for historical RSS feeds

Input: *links* is an array of URLs for RSS feeds;
start_date, *end_date*: the range over which to collect historical feed data.
Output: *DF* is a structured tabular dataset containing the publication date and title of each valid RSS feed entry.

```

1  current_date ← start_date;
2  data ← "";
3  while current_date ≤ end_date
4      for link in links
5          //send HTTP request with date parameters with exception handling:
6          fetch RSS feed for (link, current_date);
7          for entry in feed.entries
8              if publication date exists then
9                  convert publication date to datetime object;
10                 if publication_date = current_date then
11                     append entry to data;
12                 else
13                     display error message and skip entry;
14                 end if
15             end if
16         next entry
17     next link
18     increment current_date;
19 loop while
20 //convert data to a structured dataset:
21 DF ← pd.DataFrame(data);

```

```

22  if DF not empty then
23    | convert DF.date to datetime format;
24  else
25    | display error message;
26  end if
27  output DF;

```

Algorithm 5.1 was designed to collect and extract historical data from the RSS feeds of a predefined list of online news sources. The sources included major South African news platforms such as News24, Mail & Guardian, SowetanLIVE, MyBroadband, BusinessTech, Moneyweb, Eyewitness News, and IOL News. The URL links for the respective RSS feeds are tabulated below.

Table 5.11: RSS feed links of South African newspapers

South African newspaper platforms with	
Newspaper	Links
News24	https://feeds.capi24.com/v1/Search/articles/news24/TopStories/rss https://feeds.capi24.com/v1/Search/articles/news24/SouthAfrica/rss https://feeds.capi24.com/v1/Search/articles/news24/Africa/rss https://feeds.capi24.com/v1/Search/articles/news24/World/rss https://feeds.capi24.com/v1/Search/articles/fin24/tech/rss https://feeds.capi24.com/v1/Search/articles/sport/featured/topstories/rss https://feeds.capi24.com/v1/Search/articles/fin24/news/rss https://feeds.capi24.com/v1/Search/articles/channel/topstories/rss https://feeds.capi24.com/v1/Search/articles/news24/MyNews24/rss https://feeds.capi24.com/v1/Search/articles/news24/opinions/rss
Mail & Guardian	https://mg.co.za/feed/
Sowetan Live	https://www.sowetanlive.co.za/rss/?publication=sowetan-live&_gl=1*oo5jaz*_ga*MTczMDU5ODk2Ni4xNzAzOTQxMDY0*_ga_JNNJWFKJ2E*MTcwMzk0MTA2NC4xLjEuMTcwMzk0MTA4MC40NC4wLjA.ArithmeticError https://www.sowetanlive.co.za/rss/?publication=sowetan-live&section=news&_gl=1*aic8si*_ga*MTczMDU5ODk2Ni4xNzAzOTQxMDY0*_ga_JNNJWFKJ2E*MTcwMzk0MTA2NC4xLjEuMTcwMzk0MTE2MC42MC4wLjA. https://www.sowetanlive.co.za/rss/?publication=sowetan-live&section=sport&_gl=1*1s0mjr*_ga*MTczMDU5ODk2Ni4xNzAzOTQxMDY0*_ga_JNNJWFKJ2E*MTcwMzk0MTA2NC4xLjEuMTcwMzk0MTE3NS40NS4wLjA. https://www.sowetanlive.co.za/rss/?publication=sowetan-live&section=entertainment&_gl=1*9i77zi*_ga*MTczMDU5ODk2Ni4xNzAzOTQxMDY0*_ga_JNNJWFKJ2E*MTcwMzk0MTA2NC4xLjEuMTcwMzk0MTE5Mi4yOC4wLjA. https://www.sowetanlive.co.za/rss/?publication=sowetan-live&section=good-life&_gl=1*d6973n*_ga*MTczMDU5ODk2Ni4xNzAzOTQxMDY0*_ga_JNNJWFKJ2E*MTcwMzk0MTA2NC4xLjEuMTcwMzk0MTIwNC4xNi4wLjA. https://www.sowetanlive.co.za/rss/?publication=sowetan-live&section=good-life&_gl=1*d6973n*_ga*MTczMDU5ODk2Ni4xNzAzOTQxMDY0*_ga_JNNJWFKJ2E*MTcwMzk0MTA2NC4xLjEuMTcwMzk0MTIwNC4xNi4wLjA. https://www.sowetanlive.co.za/rss/?publication=sowetan-live&section=opinion&_gl=1*18ykqa8*_ga*MTczMDU5ODk2Ni4xNzAzOTQxMDY0*_ga_JNNJWFKJ2E*MTcwMzk0MTA2NC4xLjEuMTcwMzk0MTIyMS42MC4wLjA. https://www.sowetanlive.co.za/rss/?publication=sowetan-live&section=business&_gl=1*19hbel5*_ga*MTczMDU5ODk2Ni4xNzAzOTQxMDY0*_ga_JNNJWFKJ2E*MTcwMzk0MTA2NC4xLjEuMTcwMzk0MTIzNS40Ni4wLjA.

African Entertainment	https://africanentertainment.com/feed/
My Broadband	https://mybroadband.co.za/news/feed
Business Tech	https://businesstech.co.za/news/feed/
MoneyWeb	https://www.moneyweb.co.za/feed/
Eyewitness News	https://ewn.co.za/RSS%20Feeds/Latest%20News?category=Local https://ewn.co.za/RSS%20Feeds/Latest%20News?category=World https://ewn.co.za/RSS%20Feeds/Latest%20News?category=Sport https://ewn.co.za/RSS%20Feeds/Latest%20News?category=Lifestyle https://ewn.co.za/RSS%20Feeds/Latest%20News?category=Breaking https://ewn.co.za/RSS%20Feeds/Latest%20News
SA Good News	https://www.sagoodnews.co.za/feed/
Citizen	https://www.citizen.co.za/feed/
Tech Central	https://techcentral.co.za/feed/
South African Government	https://www.gov.za/feeds-view.xml https://www.gov.za/sa-feeds https://www.gov.za/news-feed
Financial Markets Journal	https://financialmarketsjournal.co.za/category/instruments-investments/feed/

The algorithm retrieved daily feeds for each day within the specified period. Data was parsed to extract key information, including the publication date, title, and summary of each entry. The algorithm ensured that only entries within the correct month were included and was stored in a structured date format. Furthermore, it handled errors like missing publication dates efficiently and provided validation in cases where no data was available.

This extraction yielded 26,964 observations recorded between September 2018 and June 2024. The data were rife with symbols, links, numbers, and noise. Furthermore, there were many days and months with no feeds, which caused a considerable number of gaps in the observations. Although the data provided a substantial amount of text, not all of them covered entire monthly periods. Fortunately, the absence of date information did not prevent observations from being used to train a machine learning model (Belanger *et al.*, 2022). Therefore, upon pre-processing and data preparation, historical RSS feeds were sufficient for developing a machine learning model for sentiment analysis. A sample of the data extracted from RSS feeds, is shown in Table 5.12.

The table illustrates the structure and format of the news data used in this study. Each entry consisted of a timestamped headline, reflecting the variety of topics, entities, and sentiment-laden language typically found in media coverage. These headlines served as the primary text input for subsequent sentiment analysis and emotion detection.

Table 5.12: Sample data extracted from RSS feeds

RSS feed data	
Date	Title
2016-01-01 00:00:00	#Mashaba100Days: 5 changes Mayor Herman Mashaba has planned for Johannesburg
2017-01-05 00:00:00	Absa calls public protector a loophole-seeking plagiarist
2017-01-18 00:00:00	Africa for optimists: 2017 in review
2017-01-26 00:00:00	Africa for pessimists: 2017 in review
2019-01-01 00:00:00	Analysis of inequality in South Africa remains shallow
2017-01-18 00:00:00	ANC membership: Anarchy on the rise
2017-01-12 00:00:00	ANC vote buying: Mbalula seizes R2.5-million
2017-01-19 00:00:00	ANC Women's League believes men used Nkosazana Dlamini-Zuma
2016-01-23 00:00:00	Black Portraits: Some that missed the mark
2018-01-31 00:00:00	British Airways launches direct flights between Durban and London

Although the sample spanned several years, only headlines within the selected study period were retained for the modelling process. The simplicity and consistency of RSS feeds made them a valuable data source for large-scale textual analysis, offering structured and time-relevant content suitable for machine learning applications. However, a notable limitation of the RSS dataset was its inability to provide data for all twelve months of the year.

Although specific dates were not required to train the model, they were essential for determining the relationship between market sentiment and stock price movements through correlation, lag and position analysis (Gauthier *et al.*, 2021). As a result, the collected RSS feed data was used for training during the initial model selection process. Model selection was an exploratory phase during which pre-processing strategies were tested and model architectures were compared to determine how they handled real-world imperfections (Nazir *et al.*, 2022).

The structure of each news website selected for data mining varied, resulting in the development of a different individual algorithm for each site. However, the basic web mining approach follows in Algorithm 5.2.

Algorithm 5.2: Web mining historical news

Input: *base_url* is the root URL of the news archive;

max_pages is the maximum number of paginated result pages.

Output: *DF* is a dataset containing the publication date and title of each article entry.

```

1  start and initialise new session;
2  articles_data  $\leftarrow$  empty list;           //set up a container to collect the results
3  for page_num  $\leftarrow$  1 to max_pages do    //repeat for a predefined number of pages
4      construct url  $\leftarrow$  base_url + page_num;    //construct the URL to visit
5      send HTTP request;
6      extract all article_containers;
7      for article in article_containers do        //repeat for all articles found
8          extract article title;
9          extract publication date;
10         if publication date is in 2018 then
```

```

11 | | append title and date to //add title and date to the data container
12 | | articles_data;
13 | | end if
14 | | delay  $\leftarrow$  2; //wait to avoid server overload
15 | end for
16 end for
17 DF  $\leftarrow$  pd.DataFrame(articles_data); //write the data to DF
18 output DF;

```

The algorithm was structured to visit multiple pages from a website constructed by the URL, which was received as input. It simulated a robot acting like a regular web browser that searched through a predetermined number of pages (input). Requests were made across multiple pages, handling pagination and dynamically parsing the HTML content. A short time delay was introduced between requests to avoid server overload. News articles were filtered by publication date, ensuring that only content from January 2018 to June 2024 was included.

The gathered data was stored in a structured DataFrame¹⁰ which enabled further analysis and exploration. Depending on the layout and structure of each website, either a set number of pages or individual articles were collected.

The web mining approach resulted in 48,896 observations that encompassed all months from January 2018 to June 2024. An excerpt from the resulting DataFrame is shown in Table 5.13.

The focus of this approach was on publicly available headlines, which provided concise representations of national discourse across various sectors like finance, politics, technology, and public affairs. These headlines formed the basis for sentiment analysis and emotion classification. Consequently, it enabled the identification of sentiment trends that may correlate with market behaviour. This structured approach ensured that the extracted content was both relevant and representative of widespread public exposure to news, supporting the development of data-driven trading strategies.

¹⁰ A DataFrame is a table-like structure resembling an Excel spreadsheet used in Python.

Table 5.13: Sample: data extracted using the web mining approach

Web mining data	
Date	Title
10-Jan-18	Kodak launches cryptocurrency
20-Jan-18	Most downloaded games on Nintendo Switch
25-Jan-18	4 ways to strengthen your organisation through cloud archiving
28-Jan-18	Bots that talk more like people
13-Feb-18	Telegram zero-day vulnerability used in cryptocurrency mining attacks
15-Feb-18	European authorities investigate Austrian Bitcoin scam
28-Feb-18	Redstor launches seamless, automated cloud backup and archiving
01-Mar-18	You should have seen that BTC Global was a scam from a mile away
23-Mar-18	SABC to investigate claims that Guptas acquired archive footage
23-Mar-18	SABC COO meeting unions unhappy with plan to commercialise SAfm

The data collected from RSS feeds served as the basis for creating and verifying the machine learning model, which is discussed in Chapter 6. While suitable for model development, the RSS dataset did not provide coverage for all twelve months of each year. This limitation was not critical for training, validating and testing the sentiment analysis model, since the development phase did not require complete temporal continuity. By contrast, the data collected through web mining included complete date fields and more comprehensive temporal coverage, which made it appropriate for time-based analyses such as correlation testing, lag analysis, and price movement analysis, that are presented in Chapter 7. In this design, the RSS feed dataset functioned as the development dataset, whereas the web mining dataset functioned as the application dataset, ensuring that the model generalised well across different data sources. The relevant testing data were used for retrospective analysis on the trading strategy covering the period from January 2024 to June 2024.

5.3 Data pre-processing

Data preparation is crucial for sentiment analysis because it directly affects the accuracy and reliability of a model's predictions (Bai *et al.*, 2020). Preparing social media data for sentiment analysis involves several steps such as pre-processing, feature extraction, and label assignment. Pre-processing refers to the removal of noise, ensuring that the data is consistent. The format of collected data determines the approach to follow in preparing it. Reformatting the stock price data according to specific rules is required, while the social media data has to be cleaned in preparation for sentiment analysis. These approaches are discussed next.

5.3.1 Stock price data

The stock price data includes the opening and closing positions, which correspond to the month-end opening and closing prices, respectively. These positions are transformed into movements

and movement percentages to analyse the potential impact of sentiment on stock price changes. To analyse monthly price dynamics, transformations are performed based on the rules that follow.

The *opening price movement* for each month was calculated as the difference between the current month's opening price and that of the previous month, denoted as

$$\Delta O_t = O_t - O_{t-1}, \quad (5.1)$$

where O_t and O_{t-1} represented the opening prices for the months t and $t - 1$, respectively. Similarly, the *closing price movement* was defined as the change in closing prices from one month to the next, expressed as

$$\Delta C_t = C_t - C_{t-1}. \quad (5.2)$$

To provide a relative measure of price fluctuation, percentage changes were also computed. The *percentage change in opening prices* was determined using the formula

$$\Delta O_t = \frac{O_t - O_{t-1}}{O_{t-1}} \times 100, \quad (5.3)$$

which quantified the monthly increase or decrease in the opening prices as a proportion of the previous month's value. Lastly, the *percentage change in closing prices* was calculated using

$$\Delta C_t = \frac{C_t - C_{t-1}}{C_{t-1}} \times 100, \quad (5.4)$$

which provided insight into the relative variation in closing prices over time. The transformations created six key variables. An example is shown in Table 5.14.

Table 5.14: Example of transformed stock price data

Date	Statistics					
	Open (R)	Open Change (R)	Open Change (%)	Close (R)	Close Change (R)	Close Change (%)
31-Jan-23	20000	610	3.15%	19845	-40	-0.20%
28-Feb-23	19547	-453	-2.27%	18178	-1667	-8.40%
31-Mar-23	18218	-1329	-6.80%	17764	-414	-2.28%
30-Apr-23	17764	-454	-2.49%	15383	-2381	-13.40%
31-May-23	15545	-2219	-12.49%	16791	1408	9.15%
30-Jun-23	16762	1217	7.83%	18945	2154	12.83%
31-Jul-23	18945	2184	13.03%	18229	-716	-3.78%
31-Aug-23	18499	-446	-2.35%	17483	-746	-4.09%
30-Sep-23	17483	-1016	-5.49%	16992	-491	-2.81%
31-Oct-23	17430	-53	-0.30%	17376	384	2.26%
30-Nov-23	17527	97	0.56%	16371	-1005	-5.78%
31-Dec-23	16371	-1156	-6.60%	16380	9	0.05%

Each row represents the end-of-month data point, with the columns showing the directional movement in prices compared to those of the previous month. This format enabled quick identification of upward or downward trends and the magnitude of changes, which was essential

for time-series modelling and sentiment-based analysis. The transformation of the stock price data into movements and movement percentages effectively prepared the data for discovering patterns that may correlate with latent sentiment about news articles.

5.3.2 Social media data

Pre-processing prepared the text data for feature extraction. In feature extraction, cleaned text was transformed into numerical codes for use in machine learning models. Each piece of data was then assigned a label that represented its sentiment.

5.3.2.1 Data preparation

A comprehensive pre-processing pipeline was implemented to transform raw news headlines into a clean, standardised, and linguistically meaningful form. This process involved several sequential steps, each designed to reduce noise and enhance the quality of the resulting input for the forthcoming feature extraction and classification tasks. The techniques used include:

- All textual content was converted to lowercase to eliminate possible inconsistencies from case sensitivity and to reduce the dimensionality of the feature space. Subsequently, regular expression-based cleaning was applied to remove various forms of noise, including HTML tags, URLs, numerical digits, and special characters. This ensured that only semantically relevant content remained.
- To further standardise the textual input, emojis and emoticons were converted into their corresponding textual representations using the *emot* library. This allowed the emotional tone often conveyed by these symbols to be retained in a machine-readable format.
- Common English stop words were removed using the Natural Language Toolkit (NLTK) *stopword* corpus, with minor adjustments to retain negation words such as 'not'. Additionally, domain-specific HTML artefacts such as 'div', 'img', 'alt', and 'p' were explicitly removed to eliminate residual structural tags from web-sourced content.
- Text was tokenised into individual words using the NLTK's *word_tokenize* function, followed by part-of-speech (POS) tagging to support the next action. Lemmatisation was performed using the WordNet *lemmatiser*, where each word was reduced to its base or dictionary form based on its POS tag. This ensured that words that were semantically equivalent but with different inflectional forms were treated uniformly, improving the generalisability of the feature space.

The cleaned and lemmatised tokens were rejoined into coherent strings for the subsequent processing stages. An excerpt of the resulting data is shown in Table 5.15.

Table 5.15: Results of pre-processing

Example of pre-processed data	
Observation	Pre-processed title
1	kodak launch cryptocurrency
2	download game nintendo switch
3	way strength organisation cloud archive
4	bot talk people
5	telegram zeroday vulnerable cryptocurrency mine attack
6	european authority investigate austrian bitcoin scam
7	redstor launch seamless automate cloud backup archive
8	global scam obvious
9	sabc investigate claim gupta acquire archive footage
10	sabc coo meet union unhappy plan commercialise

All text was transformed into lowercase and stripped of punctuation, stopwords, and irrelevant characters. This simplified and standardised format ensured that the data was clean, consistent, and ready for feature extraction.

5.3.2.2 Feature extraction

Feature extraction was conducted according to the Term Frequency-Inverse Document Frequency (TF-IDF) approach. This method transformed textual input into a numerical feature matrix reflecting the relative importance of terms within documents, while penalising commonly occurring words and emphasising more distinctive terms. The *TfidfVectorizer* tool from the Python *Scikit-learn* library was employed to vectorise the pre-processed text. Table 5.16 presents a sample of the output after vectorisation, based on the results shown previously.

Table 5.16: Vectorisation of the sample text strings

Example of vectorised data	
Observation	Vectorised title
1	379 5 186
2	59 1090 647 9
3	18 2635 2 52 7972
4	14694 8576 11
5	84 4121 6543 186 68 99
6	884 7564 115 37 2494 851
7	2938 5 1453 188 52 3693 7972
8	25 851 10455
9	1726 115 12121 789 63 7972 89
10	1726 134 31 1162 29 44 306

Each word in the pre-processed titles was replaced with a unique numerical token. This transformation converted each textual data string into a structured numerical format suitable for

use in machine learning. Each headline was represented by a high-dimensional sparse vector, with each dimension corresponding to a unique token in the corpus vocabulary.

5.3.2.3 Label assignment

In supervised learning, a model learns from labelled data. A hybrid rule-based approach was followed to assign labels using two well-established sentiment analysis tools, Valence Aware Dictionary for Sentiment Reasoning (VADER) and TextBlob, as follows:

- *SentimentIntensityAnalyzer* is a VADER lexicon and rule-based tool that returns the sentiment polarity of a given text. It is designed to take a string of text as input and return a dictionary of four scores, categorised as negative, neutral, positive, and compound. This tool was used to calculate a compound sentiment score for each headline, ranging from -1 (extremely negative) to $+1$ (extremely positive), based on the aggregation of its individual lexical features.
- TextBlob is designed for quick prototyping because it uses a built-in polarity that conveys a positive or negative sentiment and a subjectivity scoring system. It was used to generate a polarity score of between -1 and $+1$.

Sentiment scores could be translated into categorical labels using a variety of thresholding strategies. Given the continuous nature of the sentiment polarity and compound scores produced by VADER and TextBlob, there was no universally accepted standard for mapping these scores to discrete sentiment categories. As such, seven distinct labelling scenarios were selected from the literature and evaluated during model development (Ajdari & Hoti, 2023; Bethel *et al.*, 2025; Cortis & Davis, 2021; Gagiano & Marivate, 2023; Maram *et al.*, 2020; Mostofi & Shafiezed, 2024; Shabrina & Qi, 2023). These are shown in Table 5.17.

Table 5.17: Threshold configurations for sentiment label assignment scenarios

Scenario	Threshold for each sentiment, where s is the sentiment score				
	Highly positive	Positive	Neutral	Negative	Highly negative
A	-	$s > 0$	$s = 0$	$s < 0$	-
B	-	$s > 0,05$	$-0,05 \leq s \leq 0,05$	$s < -0,05$	-
C	-	$s > 0,25$	$-0,25 \leq s \leq 0,25$	$s < -0,25$	-
D	-	$s > 0,5$	$-0,5 \leq s \leq 0,5$	$s < -0,5$	-
E	$s > 0,05$	$0 < s \leq 0,05$	$s = 0$	$0 > s \geq -0,05$	$s < -0,05$
F	$s > 0,25$	$0 < s \leq 0,25$	$s = 0$	$0 > s \geq -0,25$	$s < -0,25$
G	$s > 0,5$	$0 < s \leq 0,5$	$s = 0$	$0 > s \geq -0,5$	$s < -0,5$

These scenarios were selected to assess the model's sensitivity to varying definitions of sentiment polarity and to identify the labelling strategy that yielded the most robust and generalisable performance across the dataset. The various scenarios resulted in distinct class distributions, which are discussed in Chapter 6.

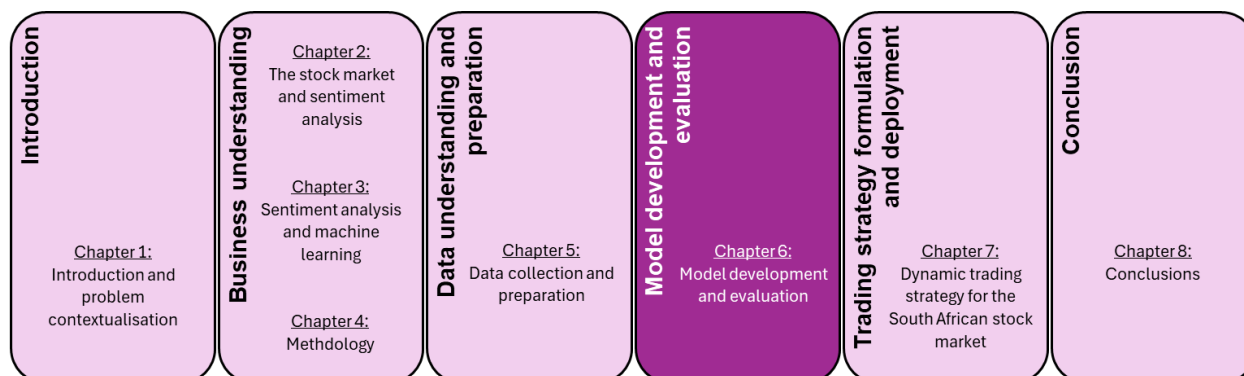
5.4 Conclusion

In this chapter, the processes of data collection, pre-processing, and preparation for sentiment analysis and stock market modelling were discussed in detail. The careful selection of stocks ensured that the study focused on companies that were representative of the South African stock market. Furthermore, the importance of high-quality data in improving the accuracy of sentiment analysis and stock market predictions was emphasised. Pre-processing techniques such as tokenisation, stop-word removal, lemmatisation, and the handling of noise were also described, as they are essential for transforming raw text into structured inputs suitable for modelling. These steps in data preparation and pre-processing form the foundation for the successful development of machine learning models.

In Chapter 6, the method of training and validating several machine learning models, to identify the most effective sentiment threshold and algorithmic configuration, is presented. This involved a systematic cross-validation process, detailed performance comparisons, and careful optimisation of hyperparameters to enhance generalisability.

CHAPTER 6

Model development and evaluation



6.1	Introduction	137
6.2	Model selection	138
6.2.1	Threshold analysis.....	138
6.2.2	Selection of models for further analysis	142
6.3	Model development.....	143
6.3.1	Impact on sentiment distribution	143
6.3.2	Hyperparameter optimisation.....	144
6.4	Final model	167
6.5	Conclusion.....	169

6.1 Introduction

In supervised learning, data is typically divided into the training, validation, and testing subsets. The training set was used to fit the model, allowing it to learn the underlying patterns in the data. The validation set was used to tune model configurations, compare different algorithms, and guide the selection of pre-processing techniques, ensuring unbiased final evaluation. The test set was held back until the end of the study to provide an independent assessment of how well the chosen model generalised to unseen data. The testing stage follows in Chapter 6.

The focus of this chapter is on the training and validation stages. These were combined in a k -fold cross-validation approach, where the data was iteratively split into smaller training and validation folds to obtain a robust estimate of model performance. These sets were used to determine the best three sentiment threshold/machine learning model combinations. A

description of this process follows in Section 6.2. The models were further evaluated based on their ability to accurately predict sentiment, while considering factors such as class imbalance and the impact of pre-processing techniques. In Section 6.3, the sentiment distribution, regularisation, and hyperparameter optimisation of the selected models are discussed. Finally, the reasoning and designation of the best pre-processing tool and overall performing model are presented in Section 6.4.

6.2 Model selection

The choice of splitting data into training and validation sets is an important methodological decision in machine learning, as it directly impacts both model learning and the reliability of performance evaluation. While common practice often defaults to ratios such as 80/20 or even 90/10, a good split may vary depending on dataset characteristics and research objectives.

For evaluating the different thresholding scenarios in this study, a 75/25 split was selected for two key reasons. First, research has shown that when data is imbalanced, a larger validation set helps ensure that the minority (scarce) classes are adequately represented, providing a more reliable estimate of model performance (Goodacre & Xu, 2018). Second, when the objective is to compare multiple machine learning models, a larger validation set improves the stability and reliability of performance comparisons by reducing variance in the evaluation metric (Einstein *et al.*, 2023).

To guide the selection of sentiment analysis tools, an exploratory investigation was conducted in which the tools TextBlob and Valence Aware Dictionary for Sentiment Reasoning (VADER) were evaluated across seven thresholding scenarios. Their outputs were used as features for a range of machine learning classifiers, namely k-nearest neighbour, logistic regression, Long Short-Term Memory (LSTM), Naïve Bayes, support vector machine, and random forests. The evaluation focused on a combined assessment of overall accuracy and the distribution of predictions across sentiment classes, ensuring that the models not only achieved acceptable predictive performance but also avoided bias towards majority classes.

6.2.1 Threshold analysis

An imbalanced label distribution occurs when the number of reference samples differs significantly across classes. Imbalanced training data can heavily influence classifier algorithms (Adam *et al.*, 2023). Moreover, each pre-processing technique has a unique impact on the accuracy and effectiveness of sentiment analysis (Garg & Sharma, 2022). The first step in the preliminary analysis was to eliminate low-performing models and those trained on severely imbalanced data, since such models are usually prone to bias toward the majority class and

unreliable generalisation. The high-level performance and class distribution of each scenario across the different pre-processing methods and label assignment techniques are discussed.

6.2.1.1 Pre-processing techniques

To enable a systematic evaluation of how the techniques and thresholds influenced the performance, a comparative analysis of the prediction accuracy for the selected machine learning models was conducted across multiple pre-processing techniques and thresholding scenarios. The techniques include lowercasing, noise removal, emojis and emoticons, stopword removal, white noise removal, and tokenisation and lemmatisation. Additionally, seven thresholding scenarios, shown in Figure 6.1, were applied to observe how variations can influence model performance.

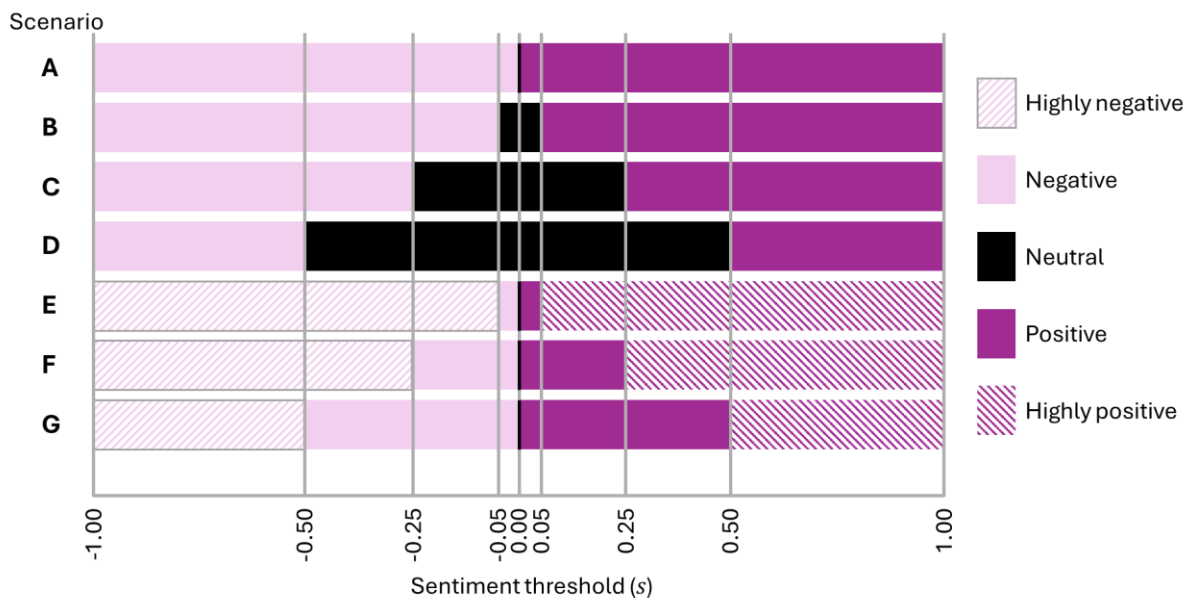


Figure 6.1: Visual representation of the sentiment thresholds considered

Each model was assessed using sentiment scores derived using TextBlob (denoted as T) and VADER (denoted as V), providing a dual perspective on the sentiment classification performance. The prediction accuracy achieved using the different pre-processing techniques is presented in Table 6.1.

Across all models evaluated, TextBlob sentiment scores consistently outperformed VADER, with the highest accuracies observed under scenario D when advanced pre-processing techniques such as tokenisation and lemmatisation were applied.

Table 6.1: Prediction accuracy (%)

Tool:	Thresholding scenario									
	A		B		C		D		E	
	T	V	T	V	T	V	T	V	T	V
K-nearest neighbour										
Lowercasing	73	65	74	65	87	69	96	85	71	64
Noise removal	76	48	67	48	86	59	96	83	64	48
Emojis and emoticons	64	48	67	48	86	59	96	83	64	48
Stopword removal	65	48	67	49	86	59	96	83	64	48
White noise removal	65	48	67	49	86	59	96	83	64	48
Tokenise and lemmatise	65	49	68	50	86	60	96	83	65	49
Logistic regression										
Lowercasing	91	85	91	85	92	85	97	88	89	82
Noise removal	92	85	91	85	92	85	97	89	89	83
Emojis and emoticons	92	85	91	85	92	85	97	89	89	83
Stopword removal	92	86	91	85	93	86	97	89	90	84
White noise removal	92	86	91	85	93	86	97	89	90	84
Tokenise and lemmatise	94	90	93	90	94	89	97	87	91	88
Naïve Bayes										
Lowercasing	75	76	72	75	85	66	96	83	73	76
Noise removal	76	77	72	75	85	66	96	83	74	76
Emojis and emoticons	76	77	72	75	85	66	96	83	74	76
Stopword removal	77	79	74	77	85	67	96	83	76	78
White noise removal	77	79	74	77	85	67	96	83	76	78
Tokenise and lemmatise	79	83	75	81	85	68	96	83	77	82
Random forests										
Lowercasing	91	81	90	80	92	81	97	89	88	78
Noise removal	91	81	90	81	92	81	97	89	88	77
Emojis and emoticons	91	81	90	81	92	81	97	89	88	77
Stopword removal	94	86	93	86	93	86	97	91	92	85
White noise removal	94	86	93	86	93	86	97	91	92	85
Tokenise and lemmatise	95	91	94	90	94	89	98	91	93	89
Support vector machine										
Lowercasing	90	84	90	84	92	84	97	91	88	81
Noise removal	91	84	90	84	92	84	97	91	88	82
Emojis and emoticons	91	84	90	84	92	84	97	91	88	82
Stopword removal	91	85	93	85	93	85	97	91	92	82
White noise removal	91	85	93	85	93	85	97	91	92	82
Tokenise and lemmatise	93	89	94	89	94	89	98	92	93	86
LSTM										
Lowercasing	94	91	93	87	93	90	97	91	92	88
Noise removal	94	91	94	91	94	92	97	91	93	91
Emojis and emoticons	95	91	94	91	94	92	98	92	93	90
Stopword removal	95	91	94	91	94	92	98	92	93	90
White noise removal	95	94	93	92	94	92	97	92	92	89
Tokenise and lemmatise	97	94	96	94	95	94	98	94	95	93

Traditional models like k-nearest neighbour and Naïve Bayes showed competitive performance, particularly when paired with TextBlob, though they were more sensitive to thresholding scenarios and pre-processing variations. Logistic regression demonstrated strong and stable results across scenarios, while random forest and support vector machine both achieved peak accuracies of 98% with TextBlob, highlighting their robustness when combined with effective pre-processing. The LSTM model emerged as the most effective overall, attaining accuracies of up to 98% and showing consistent gains from pre-processing enhancements.

Collectively, these findings prove that pre-processing quality and the choice of sentiment analysis tool play a critical role in model performance. In particular, consistent tokenisation, lemmatisation,

and noise removal reduce variance in feature representations. Accordingly, rigorous pre-processing protocols and careful tool calibration and validation are prerequisites for reproducible, high-quality results

6.2.1.2 Class distribution

To better understand the underlying sentiment structure of the dataset, the class distributions across the seven thresholding scenarios were examined. These distributions should illustrate how sentiment categories were represented within each scenario. Evaluating class distributions was essential for identifying imbalances in the sentiment categories themselves, which could significantly influence the performance of machine learning models. The distribution of sentiment classes across the seven scenarios is shown in Table 6.2.

Table 6.2: Class distributions across thresholding scenarios

TextBlob:		Scenario					
Sentiment	A	B	C	D	E	F	G
Highly positive	-	-	-	-	939	1945	355
Positive	7045	6690	3027	939	6106	4018	6690
Neutral	16599	17304	22562	25694	16599	16599	16599
Negative	3313	2963	1368	324	2989	3027	2963
Highly negative	-	-	-	-	324	1368	350
VADER:		Scenario					
Sentiment	A	B	C	D	E	F	G
Highly positive	-	-	-	-	2709	1852	355
Positive	8103	7942	6251	939	5394	6251	6690
Neutral	12059	12287	15504	25694	12059	12059	16599
Negative	6795	6728	5202	324	4566	5202	2963
Highly negative	-	-	-	-	2229	1593	350

In general, the neutral class dominated in all scenarios, but the extent of this dominance differed between the two sentiment analysis tools. TextBlob produced particularly high levels of neutrality in Scenarios C and D, which substantially reduced the number of positive and negative cases. In contrast, VADER produced a more even distribution in the earlier scenarios, showing greater sensitivity to polarity than TextBlob. This means that VADER was better at detecting subtle differences in sentiment orientation.

Another distinction lies in the emergence of extreme sentiment categories. Both tools only assigned highly positive and highly negative classes in Scenarios E, F, and G, but VADER attributed a larger share of cases to these extremes. In Scenario F, TextBlob showed a relatively balanced distribution between neutral and other sentiment classes, while VADER maintained stronger polarisation in Scenarios E and F before gradually converging toward a similar distribution pattern of TextBlob in Scenario G.

These distributions illustrated the varying degrees of class imbalance across scenarios and sentiment tools. While TextBlob tended to classify most data as neutral, VADER spread its classifications more evenly between the positive, negative, and neutral categories, particularly in the early scenarios. In contrast, VADER produced a more even distribution in the earlier scenarios, showing greater sensitivity to polarity than TextBlob. Accordingly, corrective strategies such as class weighting or resampling had to be applied during model training to prevent bias toward majority classes.

6.2.2 Selection of models for further analysis

To select three models that were to be further evaluated, some were eliminated based on the imbalanced data class distribution, which was quantified using the imbalance ratio (Choudhry *et al.*, 2023). The imbalance ratio quantifies the degree of class imbalance in a dataset and is defined as

$$IR = \frac{\max_i\{|C_i|\}}{\min_i\{|C_i|\}}, \quad (6.1)$$

where C_i denotes the number of observations in class i , $\max_i\{|C_i|\}$ is the size of the largest class and $\min_i\{|C_i|\}$ is the size of the smallest class. A higher imbalance ratio value indicates a greater disparity between the majority and minority classes, which could adversely affect model performance by biasing predictions toward the majority class. An imbalance ratio close to 1 suggests that the classes are balanced.

After eliminating models based on imbalanced class distributions, only Scenarios A and B using VADER, remained. However, Scenario A was slightly more balanced than Scenario B and was, therefore, considered the best option for further evaluation and inclusion into the trading strategy. The imbalance ratios for the positive and negative classes were 1.55 and 1.83, respectively. While the class distribution did not indicate a severe imbalance, it would be crucial to closely monitor performance metrics such as precision, recall, and F1-scores for the minority classes. These refer to those sentiment categories with significantly fewer observations when compared to the majority class.

Understanding the degree of class imbalance through this ratio provided an important context for model training, as it would influence the choice of algorithms, evaluation metrics, and potential resampling techniques. With this foundation, the focus of the following section is on model development, during which appropriate strategies were applied to ensure that the predictive models performed reliably across all sentiment classes.

The initial pool of machine learning models was selected based on their frequency of use and demonstrated effectiveness in prior financial sentiment analysis studies. Rather than

implementing an exhaustive set of classifiers, the study focused on representative algorithms consistently cited in the literature. Model selection for final evaluation was based on balanced performance across sentiment classes, robustness to preprocessing variations, and stability across validation periods. Therefore, the logistic regression, random forests, and LSTM were selected for further development due to their complementary strengths in handling skewed class distributions. This ensured methodological alignment with existing research while maintaining empirical rigour and the predictive models performed reliably across all sentiment classes while capturing both classical and sequential patterns in the text data.

6.3 Model development

In this section, the refined evaluation of the three selected models is discussed, with a focus on identifying signs of underfitting and overfitting. A 10-fold cross-validation approach was employed by dividing the data into different 75/25 training/validation splits. This iterative process ensured that every observation in the dataset was used for both training and validation at different stages, thereby reducing bias introduced by any single split. This strategy not only provided a more robust estimate of model performance across different thresholding scenarios but also enhanced the reliability of model selection by averaging results over multiple folds. The training and validation accuracy, as well as the training and validation loss, were monitored and evaluated over time to assess the models' learning behaviour and convergence.

6.3.1 Impact on sentiment distribution

Both logistic regression and random forest were adjusted for class imbalance using the class weight parameter. It controls per-class penalties during training and includes options *balanced* or *None*. Logistic regression is especially prone to bias toward the majority class in multiclass settings (Geisler *et al.*, 2022), while random forests handle imbalance more robustly but still benefit from re-weighting (Budi *et al.*, 2020). To address this imbalance, the class weight parameter was set to *balanced* so that each class weight is inversely proportional to its observed frequency, which reduces majority-class bias and improves the classification of under-represented sentiment categories.

LSTM is susceptible to issues caused by sentiment class-imbalanced data (Giang *et al.*, 2018). Oversampling was applied to the dataset to address the issue of class imbalance. It is a class-imbalance remedy applied only to the training data that increases the number of minority-class examples so that the classifier sees a more balanced class distribution. The process involved balancing the class distribution by increasing the representation of the minority classes. First, the class with the highest number of instances in the target column was identified as the dominant class. For each sentiment class, the dataset was filtered, and resampling with replacement was

used to increase the sample size of the minority classes to match the count of the dominant class. This ensured that all classes had an equal number of samples, reducing the impact of class imbalance. The oversampled datasets were then concatenated to form a balanced dataset for training the model. The process was controlled by setting a random seed¹¹ to ensure reproducibility. The imbalance ratio was now equal to one for all classes.

Effective model performance in machine learning is not solely dependent on the choice of algorithm, but also on the selection of hyperparameter settings and tuning. Hyperparameter optimisation ensures that a model is configured to achieve its best possible predictive capacity while avoiding underfitting or overfitting. Regularisation techniques further contribute by controlling model complexity and enhancing generalisation to unseen data.

6.3.2 Hyperparameter optimisation

In this section, the focus is on hyperparameter optimisation, and the regularisation strategies employed for the three models under investigation. Hyperparameters are parameters defined before training, such as the learning rate, number of epochs, and model architecture, which guide how the model learns (Dalal *et al.*, 2024). Regularisation is a component of hyperparameter optimisation and involves techniques that control model complexity to improve generalisation and reduce overfitting (Ho & Kee, 2025). For each model, the hyperparameter limits are shown, followed by the optimisation methods and the resulting impact on predictive performance. Generalisability refers to the model's ability to perform well on new data and is reflected in how closely the training and validation losses align (Van Den Broek & Van Der Sluis, 2025). A divergence between these losses - where the training loss continues to decrease while the validation loss begins to increase - indicates overfitting (Dishar & Muhammed, 2023). Therefore, generalisability is defined by how close the training and validation losses were, and the change in the training loss indicates overfitting, while the validation loss started to increase over time.

The model evaluation begins with an assessment of the confusion matrix, a fundamental tool used to summarise the performance of a classification model. A confusion matrix provides a detailed breakdown of how well the model's predicted classes align with the actual (true) classes by displaying the number of correct and incorrect predictions for each sentiment category. It consists of four key components: true positives (TP), where the model correctly predicts the positive class; false positives (FP), where the model incorrectly predicts a positive outcome; false negatives (FN), where the model fails to identify a true positive; and true negatives (TN), where

¹¹ A random seed is the initial value that initialises a pseudorandom number generator so the produced "random" sequence is repeatable.

the model correctly predicts the absence of the target class. Together, these values offer a comprehensive view of model accuracy, precision, recall, and F1-score. It highlights both correct classifications and misclassifications across sentiment categories. Analysing the confusion matrix provides insight into the model's strengths, weaknesses, and potential areas for improvement in sentiment prediction.

Additionally, the Area Under the Receiver Operating Characteristic (AUC-ROC) curve serves as a threshold-independent measure of the model's ability to distinguish between classes. The ROC curve plots the true positive rate against the false positive rate, and the area under this curve quantifies the overall separability between classes. A higher AUC value indicates superior discriminative capability, meaning the model is more effective at correctly identifying positive and negative sentiments across varying decision boundaries. However, excessively high AUC values on training data may indicate overfitting, where the model learns patterns too specific to the training set and fails to generalise well to unseen data. Conversely, low AUC values may suggest underfitting, where the model is too simple to capture the underlying relationships in the data. To evaluate this balance between prediction confidence and correctness, logarithmic loss is applied as an additional metric, quantifying the divergence between predicted probabilities and actual class labels to provide a more nuanced measure of model calibration.

6.3.2.1 Logistic regression

Logistic regression does not involve complex configurations, such as layers or extensive hyperparameters, resulting in a more concise configuration. The initial performance of this model is displayed in the form of the confusion matrix in Figure 6.2.

Both of the raw classification counts and the corresponding row-normalised proportions are shown in parentheses. The diagonal cells represent correct classifications, while off-diagonal cells indicate misclassifications.

Confusion matrix for the initial logistic regression

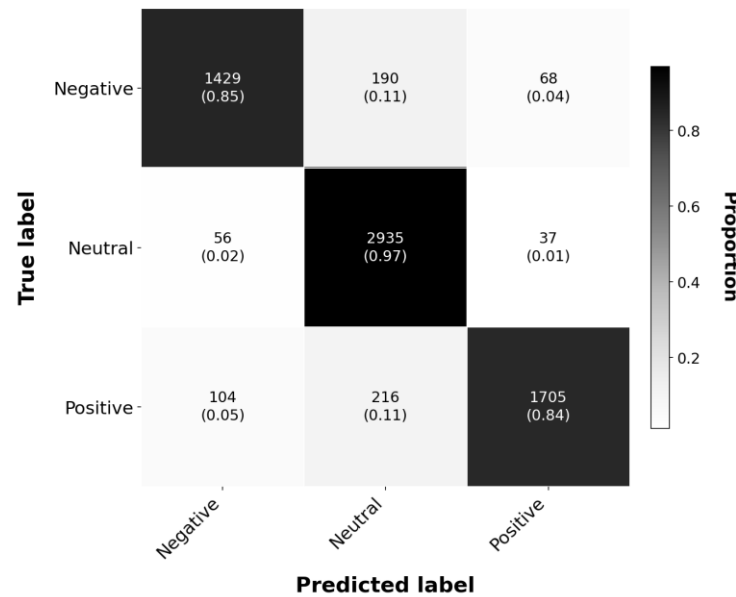


Figure 6.2: Confusion matrix for the initial logistic regression

The model demonstrated strong performance across all sentiment categories, with a particularly high accuracy for the *neutral* class (97%), followed by *negative* (85%) and *positive* (84%) sentiments. This suggested that the model effectively distinguished neutral text but exhibited a minor overlap between *positive* and *negative* predictions. This is typical in sentiment classification tasks due to contextual and linguistic ambiguity. The values in parentheses represent row-normalised proportions which indicate the proportion of predictions within each true class that fell into each predicted category. For instance, 0.85 under the negative-to-negative cell signifies that 85% of all true negative instances were correctly predicted as *negative*. This normalisation method provided an interpretable view of class-level performance, allowing proportional comparison across classes regardless of differences in class size. The confusion matrix translates to the performance metrics in Table 6.3.

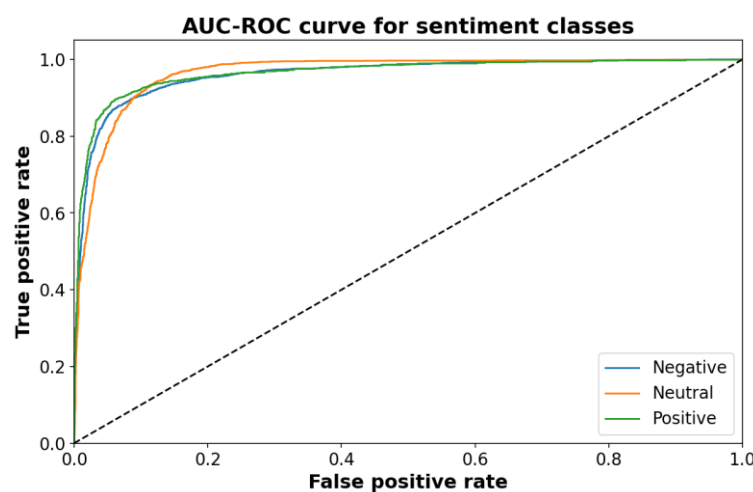
Generally, the metrics indicated that the model performed well in classifying sentiments, particularly in identifying positive and neutral sentiments. Although the overall model accuracy is 0.90, the individual sentiment accuracies differed slightly, with negative, neutral, and positive classes achieving 0.77, 0.86, and 0.80 respectively. This difference arose because overall accuracy was computed across all classifications in the confusion matrix, considering the total number of correct predictions out of all predictions made.

Table 6.3: Classification report for the initial logistic regression model

Sentiment	Accuracy	Precision	Recall	F1-score	Comment
Negative	0.77	0.90	0.85	0.87	Most negatives were correctly predicted (high TP, few FN). Some neutral and positive samples were misclassified as negative (moderate FP). Overall, balanced and reliable detection.
Neutral	0.86	0.88	0.97	0.92	Strong recall due to high TP but increased FP from negative and positive classes. Slight bias toward predicting neutrality.
Positive	0.80	0.94	0.84	0.89	High TP and precision with few FP. Some positive samples were misclassified as neutral (FN), but there was an overall strong polarity distinction.

In contrast, per-sentiment accuracy reflected only the proportion of correct predictions within each sentiment category. Performance metrics such as training and validation loss provided more insight into model behaviour.

In addition to these metrics, the AUC-ROC curve was used to evaluate the model's ability to distinguish between sentiment classes at various threshold settings. The AUC-ROC curve for the initial logistic regression model is presented in Figure 6.3.

**Figure 6.3: AUC-ROC curve for the initial logistic regression model**

All three sentiment curves are positioned above the diagonal reference line. This indicates that the model performed better than random classification. However, the consistently high AUC values across classes might suggest potential overfitting, as the model appeared to perform exceptionally well on the training data but might not generalise effectively to unseen data. This observation motivated the refinement of the logistic regression model through parameter optimisation and class balancing in subsequent iterations. Moreover, the logarithmic loss quantified the difference between the predicted probabilities and true class labels.

Hyperparameters such as regularisation strength and solver choice also affected model performance. By visualising the training and validation performance through loss curves, a model learning process could be assessed, and necessary adjustments could be made to improve its generalisation capabilities. The log loss of the logistic regression model, over 100 iterations of training, is shown in Figure 6.4.

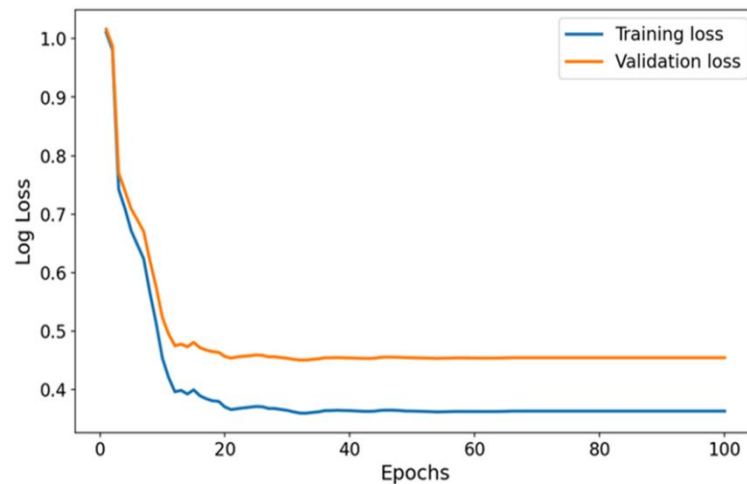


Figure 6.4: The bias-variance trade-off for validation loss of the initial logistic regression model

Both curves show a rapid decline during the first 10 to 12 epochs, indicating that the model quickly learned key patterns in the data. Beyond this point, the testing loss stabilised at an approximate loss value of 0.35, while the training loss continued to decline gradually to just below 0.30. The small but stable difference between the two curves following 0.35 suggested mild overfitting. Overall, the graph indicates that the model fitted the training data relatively well while retaining good generalisation to the validation set. Early stopping at around ten to twelve epochs could prevent unnecessary overfitting while preserving generalisation performance. The final bounds of the hyperparameters are summarised in Table 6.4.

The only parameter explicitly explored across a numeric range was the maximum number of iterations, which was varied between one and 12 to study convergence behaviour through the loss curves. Therefore, both the training and validation losses stabilised without signs of instability or underfitting. Since this configuration already produced consistent results and satisfactory generalisation, more complex hyperparameter optimisation methods were not required. Introducing additional optimisation for other parameters would likely have provided a marginal benefit while increasing computational complexity without significant performance improvement.

Table 6.4: Upper and lower bounds of hyperparameters in the revised logistic regression model

Hyperparameter	Lower bound	Upper bound	Explanation
Class weight	-	-	Fixed to <i>balanced</i> to compensate for class imbalance by weighting classes inversely to their frequencies.
Maximum iterations	1	12	Varied across this range when plotting the loss curves. This allowed observation of optimisation behaviour over increasing training cycles. The results showed that both training and testing losses stabilised after approximately 10 to 12 iterations.
Penalty type	-	-	Not applicable as regularisation techniques were not used.
Regularisation strength (C)	1	1	Fixed to 1 as no regularisation techniques were used. Therefore, no stronger or weaker penalisation was implemented.
Tolerance (stopping criterion)	1×10^{-4}	1×10^{-4}	Fixed stopping tolerance, determining when tuning halts once improvements fall below this threshold.

All other hyperparameters, such as regularisation strength and class weight, were fixed. The performance of the revised model, with early stopping applied at 12 epochs, is shown in Figure 6.5.

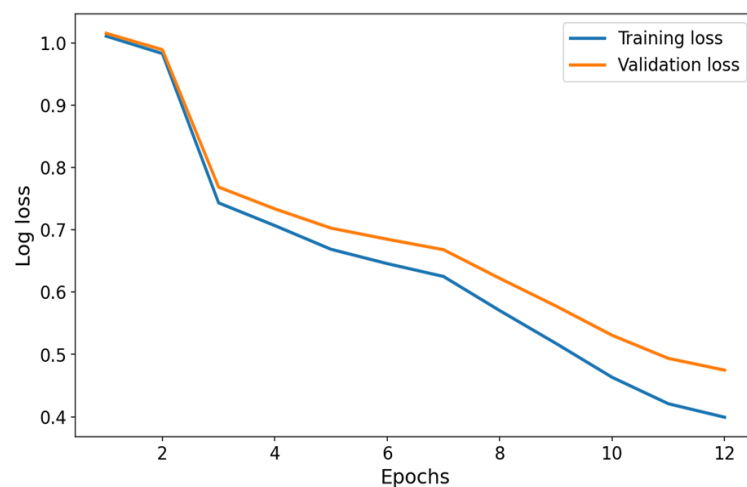


Figure 6.5: The bias-variance trade-off for validation loss of the revised logistic regression model

Both loss curves showed a steady decrease, with the steepest decline occurring during the first three epochs, indicating that the model quickly learned the key patterns in the data. Beyond this point, the training and testing losses continued to decrease at a slower and more consistent rate, reaching approximately 0.40 and 0.48, respectively, by epoch 12. These values were considered acceptable within the context of logarithmic loss, where lower values indicate better calibration between predicted probabilities and actual class labels. The small difference between training

and testing losses further suggested limited overfitting and stable generalisation across data subsets. This confirmed that the chosen early stopping-point of 12 epochs effectively balanced training performance and validation accuracy. The confusion matrix for the revised model is presented in Figure 6.6.

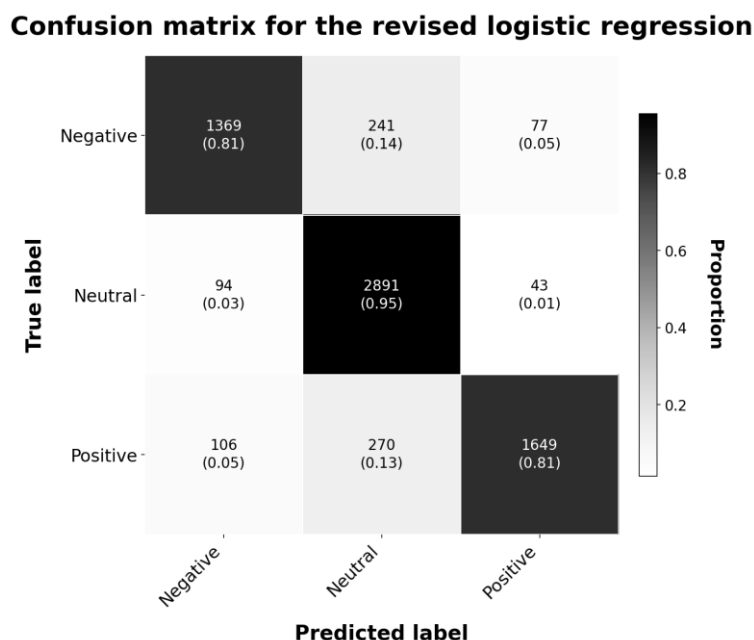


Figure 6.6: Confusion matrix for the revised logistic regression model

The majority of the predictions aligned correctly with their true labels across all sentiment categories, with the highest proportion observed for the neutral class (0.95), followed by the negative and positive classes (both 0.81). This indicated that the model effectively distinguished between sentiment categories after hyperparameter optimisation and class balancing. The reduction in false positives and false negatives across all classes suggested improved model calibration and generalisation, resulting in fewer misclassifications. The darker diagonal cells represented a strong agreement between predicted and actual labels, while the lighter off-diagonal cells reflected the limited number of incorrect predictions. Overall, the revised model demonstrated robust performance and a more balanced capability to classify sentiments accurately across the dataset. The performance measures for the revised model are summarised in Table 6.5.

Although a slight decrease was observed in the overall accuracy of 0.88 and per-sentiment accuracies, this reduction reflected an improved model generalisation rather than degraded performance. The revised model demonstrated more balanced precision, recall, and F1-scores across sentiment categories, suggesting that it captured underlying sentiment patterns better without overfitting to the training data. In practice, it is preferable to have a model with moderately

lower accuracy that fits the data well and performs reliably on unseen data than one that exhibits artificially high accuracy due to overfitting.

Table 6.5: Classification report for the revised logistic regression model

Sentiment	Metrics							
	Accuracy	Change	Precision	Change	Recall	Change	F1-score	Change
Negative	0.73	−0.04	0.84	−0.06	0.86	+0.01	0.85	−0.02
Neutral	0.82	−0.04	0.84	+0.04	0.96	−0.01	0.90	+0.02
Positive	0.77	−0.03	0.91	+0.03	0.86	−0.02	0.89	0

The adjustments made through class balancing and hyperparameter optimisation, therefore, produced a model that generalises more effectively, yielding stable and interpretable results across all sentiment classes. Consider the AUC-ROC curve for the revised logistic regression model in Figure 6.7.

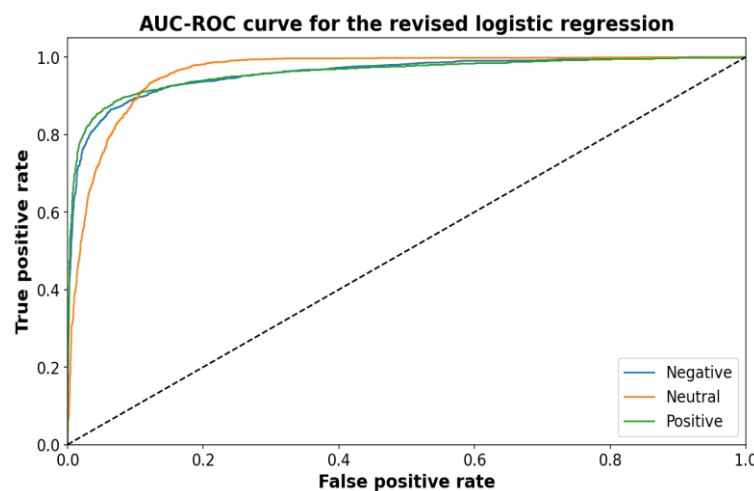


Figure 6.7: AUC-ROC curve for the revised logistic regression model

All three sentiment classes showed curves positioned well above the diagonal reference line, indicating performance superior to random classification. Compared to the initial model, the curves were smoother and more stable. This suggested improved calibration and a reduction in overfitting. Overall, the AUC-ROC results confirmed that the revised logistic regression model achieved strong and balanced class discrimination while maintaining better generalisation to unseen data.

6.3.2.2 Random forests

An ensemble learning method, the random forests model builds multiple decision trees and then aggregates their predictions to improve accuracy and control overfitting. In this study, a classifier comprising 20 trees was implemented to classify sentiment categories. The model initially used

20 trees to establish a computationally efficient baseline for comparison; starting with a smaller number of estimators allowed for faster training and quicker evaluation of the model's behaviour and generalisation ability before introducing additional complexity. Each tree in the forest was constructed from a random subset of the training data and features, introducing diversity and reducing the likelihood of overfitting to any single dataset. This method utilised the ability of decision trees to capture non-linear relationships while reducing their individual tendency to overfit by combining multiple trees into a robust ensemble.

Predictions were obtained by aggregating outputs from all trees, which reduced variance compared to individual decision trees and improved generalisation. The model's performance was initially evaluated using accuracy, precision, recall, and F1-score, while overfitting was monitored by comparing training and validation losses. The model was evaluated based on its accuracy, which measured how well it could classify unseen test data. The initial performance of this model is shown in the confusion matrix in Figure 6.8.

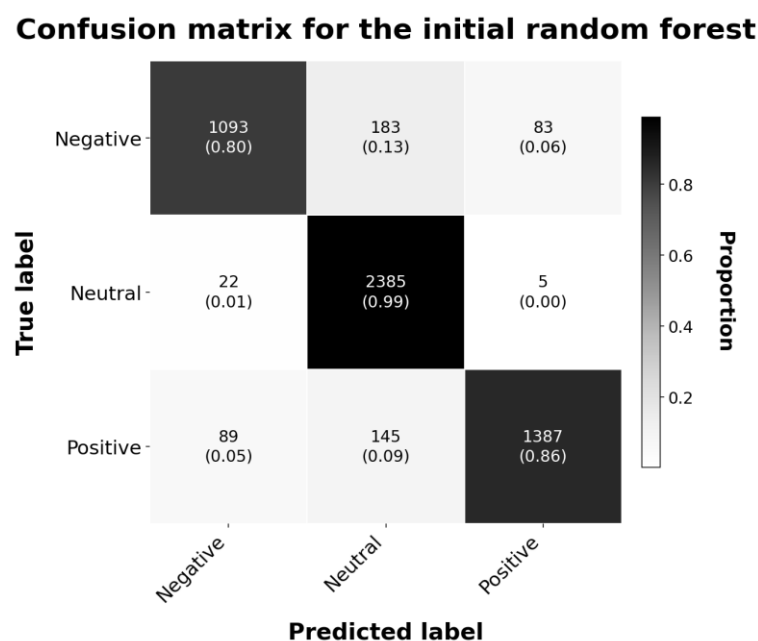


Figure 6.8: Confusion matrix for the initial random forests model

The model achieved high true positive rates across all classes, particularly for neutral sentiment, where 0.99 of samples were correctly classified. The positive and negative classes also demonstrated strong predictive performance, with 0.86 and 0.80 of the samples correctly identified, respectively. However, the dominance of the neutral class suggested that the model might be biased toward predicting the majority sentiment, a common occurrence in imbalanced datasets. This tendency could lead to overfitting, where the model learnt the patterns of the more frequent class at the expense of minority classes. The small but notable proportion of misclassifications between negative and positive sentiments indicated an overlap in linguistic

expression, which could be challenging to distinguish. Overall, while the initial random forests model achieved high accuracy, the distribution of predictions suggested the need for hyperparameter optimisation to enhance generalisation and reduce class bias. The performance metrics derived from the confusion matrix for the initial random forests model, are shown in Table 6.6.

Table 6.6: Classification report for the initial random forests model

Sentiment	Metrics				Comment
	Accuracy	Precision	Recall	F1-score	
Negative	0.74	0.91	0.80	0.85	High precision indicates few FPs, but lower recall suggests more FNs.
Neutral	0.87	0.88	0.99	0.93	Very high recall with slightly lower precision indicates some FPs but strong detection of true neutral cases.
Positive	0.81	0.94	0.86	0.90	High precision and recall show balanced performance, with few FPs and FNs.

The model's performance across the three sentiment classes demonstrated strong overall precision, recall, and F1-scores. The neutral class had the highest recall at 0.99, indicating that the model was highly effective at identifying neutral sentiments, though it achieved a slightly lower precision of 0.88. This suggested that while the model accurately predicted neutral sentiments, some of its neutral predictions might have included other classes. The negative class had a precision of 0.91, showing that it correctly identified most negative instances, though its recall of 0.80 indicated that some negative sentiments had been misclassified. Similarly, the positive class had a high precision score but a slightly lower recall of 0.84, meaning that while it accurately predicted positive sentiments, a few positive instances had been overlooked. The overall accuracy amounted to 0.90.

Overall, the model balanced precision and recall well, with F1-scores above 0.85 for all classes. This showed solid performance in sentiment classification. In addition to these metrics, the AUC-ROC curve for the initial logistic regression model is presented in Figure 6.9.

All sentiment curves lay well above the diagonal reference line, indicating that the model performed substantially better than random classification. The neutral sentiment achieved the highest AUC, reflecting its dominant representation in the dataset and the model's strong ability to identify neutral samples correctly. The positive and negative sentiment curves also demonstrated high separability, though with slightly lower AUC values due to a small proportion of misclassifications between these categories.

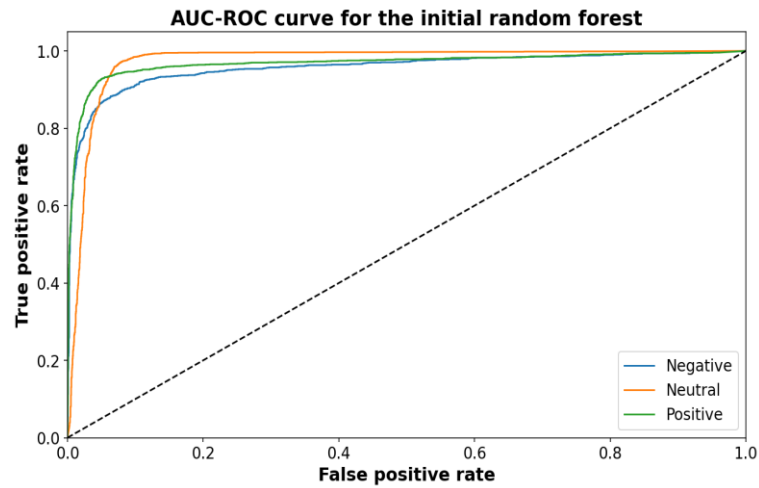


Figure 6.9: AUC-ROC curve for the initial random forests model

The consistently high AUC across all classes suggests that the initial random forests model exhibits strong discriminative capability; however, this may also indicate potential overfitting, as near-perfect separation on the training data can lead to reduced generalisation on unseen data. To address this, hyperparameter optimisation was employed to optimise the model's complexity and improve its balance between bias and variance. This process began with examining the training and validation loss curves, which provided insight into whether the model fitted too closely to the training data or failed to capture underlying patterns effectively. The graph in Figure 6.10 demonstrates the relationship between the number of trees in the random forests model and the corresponding training and validation loss.

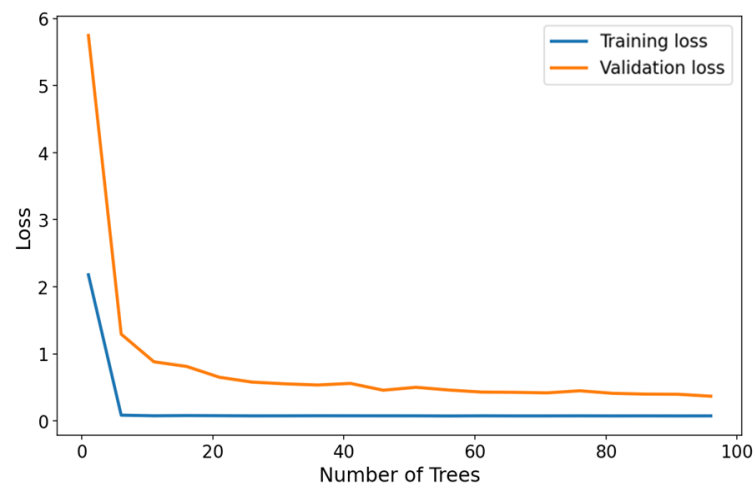


Figure 6.10: Bias-variance trade-off for validation loss of the initial random forests model

Training and validation loss both decreased sharply as the number of trees increased from one to approximately ten, after which the losses stabilised. This indicated diminishing returns in performance improvement as more trees were added. Since the training loss remained close to

zero and the validation loss stabilised without diverging, the model did not exhibit signs of overfitting or underfitting. However, tuning parameters such as the maximum tree depth or the number of features considered at each split might yield better performance and further optimise generalisation. The performance improvement became negligible at approximately 20 to 30 trees. Therefore, the number of estimators was fixed at 20 to achieve a balance between model performance and computational efficiency.

An exhaustive, or grid search¹², approach was implemented for the hyperparameter optimisation process. Key parameters such as the number of trees, maximum tree depth, and minimum samples required for splitting and leaf nodes could be varied and validated using a cross-validation approach. This process aimed to identify an optimal configuration that balanced the model complexity with performance. In this case, loss was defined as the classification error. If the training accuracy was significantly higher than the validation accuracy, or if the validation loss increased while the training loss decreased, it was a sign of overfitting. This process aimed to identify an optimal configuration that balances model complexity with performance. The bounds of the hyperparameters are summarised in Table 6.7.

The most critical variation was the number of trees, tested from one to 30, as this parameter directly determined the ensemble's ability to generalise and stabilise predictions. A random forest aggregates the outcomes of multiple decision trees to reduce variance and improve robustness. Too few trees may lead to high variance and unstable predictions, as individual trees can dominate the ensemble. At the same time, too many trees may increase computational cost without significant performance gains once convergence is reached. Therefore, optimising the number of trees was essential to balance model accuracy, variance reduction, and computational efficiency.

¹² Grid search is a technique that systematically tests all possible combinations of specified hyperparameter values to identify the set that yields the best model performance.

Table 6.7: Upper and lower bounds of hyperparameters in the revised random forests model

Hyperparameter	Lower bound	Upper bound	Explanation
Number of trees	1	30	The number of decision trees in the ensemble was explored within this range. A smaller number of trees increases bias, while a larger number improves stability but raises computational cost.
Minimum samples required to split	2	50	This parameter was varied to regulate the minimum number of observations required to split an internal node. Lower values allow highly detailed trees, whereas higher values restrict splitting and encourage simpler structures.
Minimum samples per leaf	1	50	The number of observations required at the leaf level ranged from very fine granularity (1) to more generalised groupings (50). This balance influences overfitting and generalisation.
Maximum depth of trees	5	15	The depth of the trees was limited to values between 5 and 15, providing a balance between capturing detailed interactions and preventing excessive complexity.

Split and leaf size constraints were also adjusted to regulate tree growth, while the maximum depth was explored at multiple levels (five to 15). A fixed numeric setting of 25 was additionally trialled for the maximum number of features. Together, these bounds enabled systematic evaluation of model complexity and regularisation effects. The loss performance of the revised model is shown in Figure 6.11.

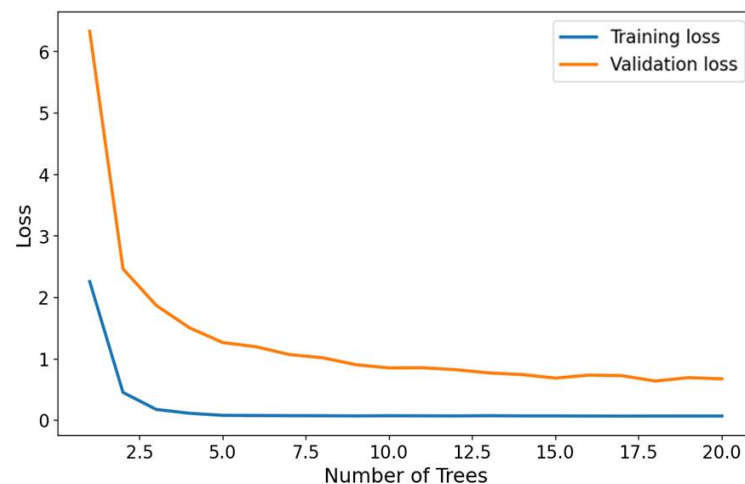


Figure 6.11: The bias-variance trade-off for validation loss of the revised random forests model

A steep reduction in both loss profiles was observed for one to three trees, demonstrating that the ensemble quickly improved its predictive performance with the addition of trees. With

approximately five trees, the training loss approached zero, as fully grown decision trees could perfectly fit the training data. The validation loss continued to decline gradually and began to stabilise with 15 or more trees, indicating diminishing returns from adding more trees. The relatively small and stable gap between training and validation loss suggested that the model generalised well. Consider the confusion matrix for the revised model in Figure 6.12.

Confusion matrix for the revised random forest

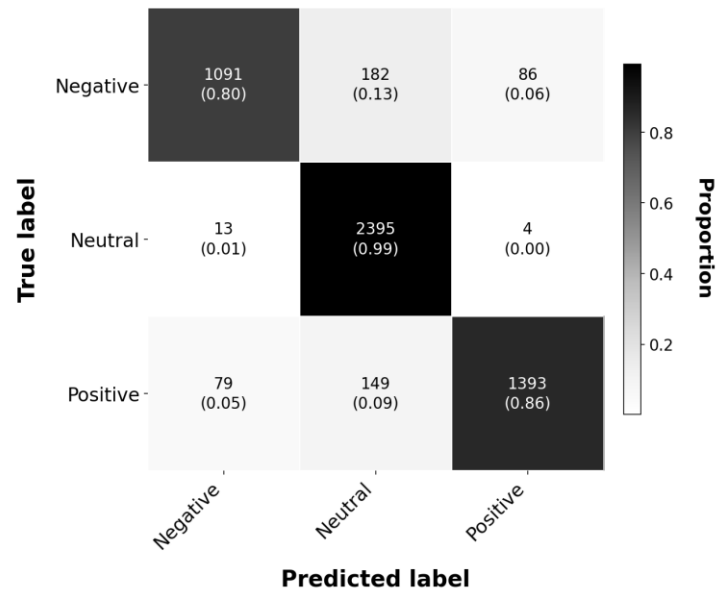


Figure 6.12: Confusion matrix for the revised random forests model

The model demonstrated consistently strong predictive performance across all sentiment classes, with the majority of samples correctly classified. The neutral class achieved a near-perfect classification accuracy, with 0.99 of neutral instances correctly identified, indicating that the model effectively captured the dominant sentiment patterns in the dataset. The negative and positive classes also achieved high true positive rates of 0.80 and 0.86, respectively, with relatively few misclassifications. The reduction in false positives and false negatives compared to the initial model suggested that hyperparameter optimisation improved the model's calibration and generalisation.

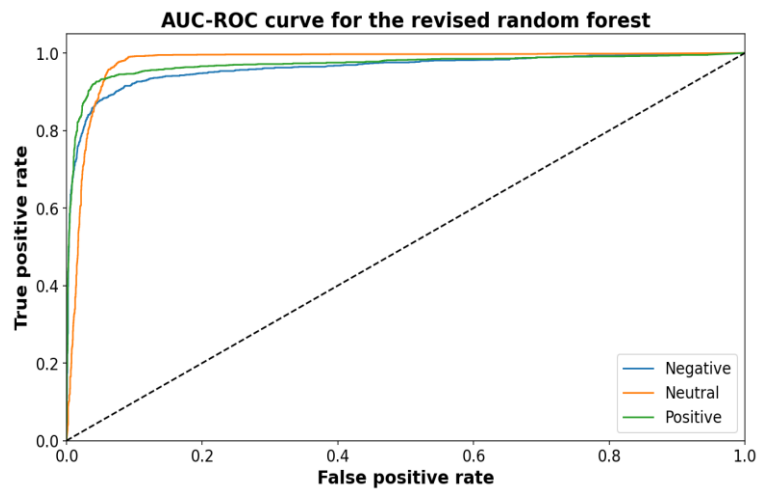
The more balanced prediction distribution across sentiment categories indicated that the model no longer overfitted the training data and better reflected the underlying structure of the text features. Overall, the revised random forests model demonstrated enhanced stability and reliability in sentiment classification performance. The performance measures for the revised model are summarised in Table 6.8.

Table 6.8: Classification report for the revised random forests model

Sentiment	Metrics							
	Accuracy	Change	Precision	Change	Recall	Change	F1-score	Change
Negative	0.75	+0.01	0.91	−0.01	0.84	+0.02	0.87	+0.01
Neutral	0.87	+0.00	0.89	+0.01	0.99	+0.00	0.93	+0.00
Positive	0.81	+0.00	0.94	+0.00	0.84	+0.00	0.89	+0.00

The results indicated marginal but meaningful improvements across most evaluation measures, particularly for the negative sentiment class, where accuracy and F1-score increased by 0.01 and 0.02 respectively. The precision and recall values remained consistently high across all classes. This demonstrated the model's stability and balanced predictive capability. The neutral sentiment maintained a near-perfect recall (0.99), which suggested that the model continued to identify the majority sentiment class effectively, without bias. The slight variations in precision and recall reflected a more generalised fit following hyperparameter optimisation, which successfully mitigated overfitting while preserving model accuracy.

Overall, the revised random forest achieved a strong equilibrium between precision, recall, and F1-score across all sentiment categories. Moreover, consider the AUC-ROC curve for the revised model in Figure 6.13.

**Figure 6.13: AUC-ROC curve for the revised random forests model**

The curves for all three classes remained consistent, indicating a strong discriminative performance and a low false positive rate. Compared to the initial model, the revised version showed smoother and more stable curve trajectories, suggesting an improved generalisation after hyperparameter optimisation. The near saturation of the true positive rate for the neutral and positive classes confirmed the model's effectiveness in correctly identifying dominant sentiment patterns, while maintaining balanced sensitivity across all classes. The high overall AUC values

reinforced/proved/endorsed that the random forests model achieved a reliable trade-off between recall and specificity, effectively capturing sentiment polarity with minimal misclassification on unseen data.

6.3.2.3 Long Short-Term Memory

An LSTM network is a type of RNN designed to learn long-range dependencies by using gated memory cells that regulate which information is retained, updated, or forgotten over time. The architecture began with an embedding layer that converted each input word into a 250-dimensional dense vector, enabling the model to capture semantic relationships and reduce input dimensionality. The embedded sequences were then processed by the first LSTM layer, followed by a dropout layer that deactivated half of the layer's units at each training step, thereby reducing the risk of overfitting. A second LSTM layer further refined the learned temporal patterns before the model produced its final prediction through a dense output layer with a *softmax* activation function. The structure of this initial model is illustrated in Figure 6.14.

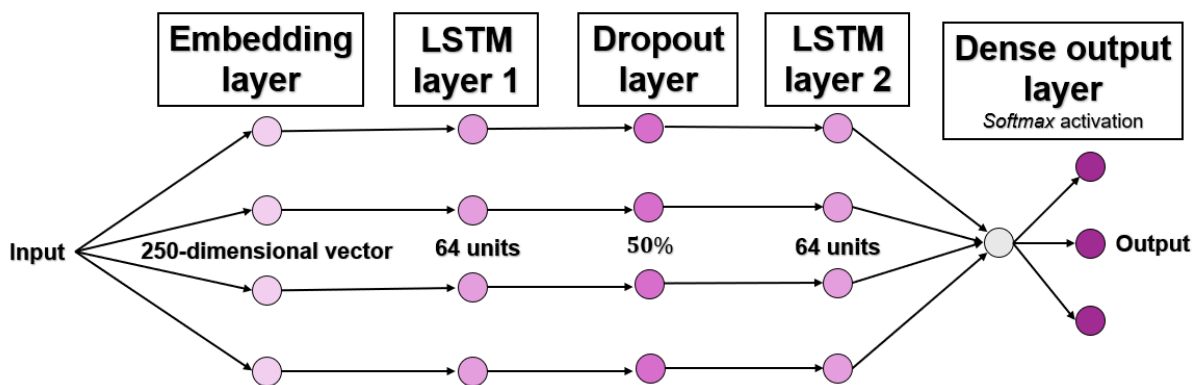


Figure 6.14: Visualisation of the initial LSTM model

The dropout layer gave output to the full sequence to capture temporal dependencies across time steps. In contrast, the last layer condensed this information into the final hidden state used for classification.

A layer with a rate of 50% dropout was applied to prevent overfitting and improve model generalisation. The dense output layer consisted of three neurons corresponding to the sentiment classes - positive, neutral, and negative - and used a *softmax* activation function to generate class probabilities. The model was optimised using the *Adam* algorithm and trained with categorical cross-entropy loss to ensure robust multi-class classification performance. The performance of this initial model is illustrated in the confusion matrix in Figure 6.15.

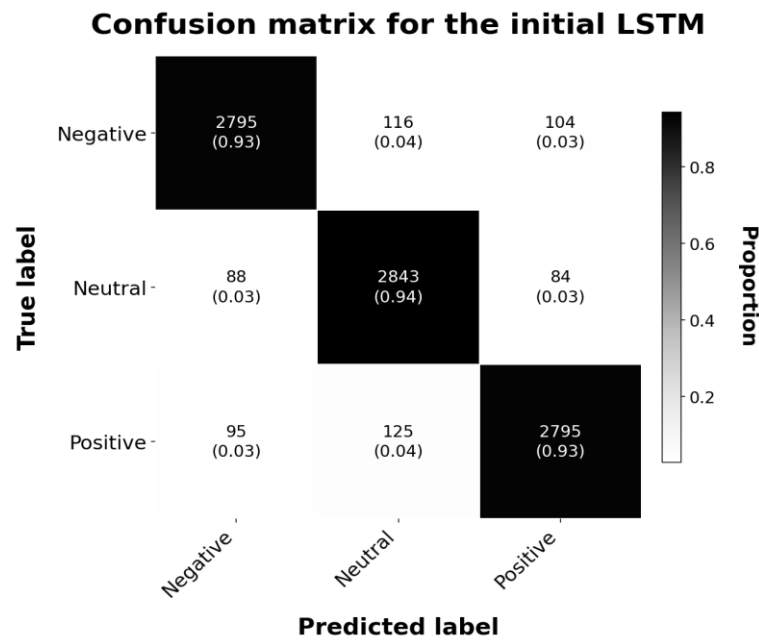


Figure 6.15: Confusion matrix for the initial LSTM model

The confusion matrix for the initial LSTM model shows a strong classification performance across all three sentiment categories. The majority of predictions fell along the diagonal, which represented correctly classified instances, with proportions of 0.93 for both negative and positive sentiments, and 0.94 for neutral sentiment. This indicated that the model effectively learned the temporal dependencies within the data, enabling accurate sentiment differentiation. The small off-diagonal values (ranging between 0.03 and 0.04) represented limited misclassifications between sentiment classes. It suggested a strong generalisation and low overlap between classes.

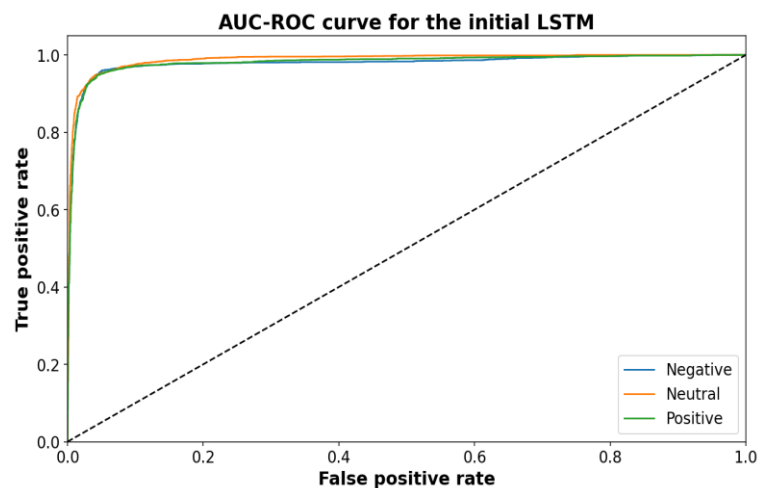
These results showed that the LSTM architecture successfully captured long-term contextual patterns in the text, leading to consistent and well-balanced sentiment predictions. However, this level of uniform accuracy across sentiment classes may suggest mild overfitting. The performance metrics derived from the confusion matrix for the initial LSTM model are shown in Table 6.9.

Sentiment classification results reflected strong predictive performance, reliable accuracy and consistency across all sentiment classes. Furthermore, the model achieved an overall accuracy of 0.93. For further investigation, the AUC-ROC curve for the initial logistic regression model is presented in Figure 6.16.

Table 6.9: Classification report for the initial LSTM model

Sentiment	Metrics				Comment
	Accuracy	Precision	Recall	F1-score	
Negative	0.87	0.94	0.93	0.93	Balanced classification with low FPs and FNs.
Neutral	0.87	0.92	0.94	0.93	Slightly higher recall indicates effective detection of neutral sentiments, though minor FPs may occur.
Positive	0.87	0.94	0.83	0.93	High precision but lower recall, suggesting some positive samples were misclassified as other sentiments.

Sentiment classification results reflected strong predictive performance, reliable accuracy and consistency across all sentiment classes. Furthermore, the model achieved an overall accuracy of 0.93. For further investigation, the AUC-ROC curve for the initial logistic regression model is presented in Figure 6.16.

**Figure 6.16: AUC-ROC curve for the initial LSTM model**

The curves for all three sentiment classes remained close to the top-left corner which indicates strong separability and a high overall AUC value. They demonstrated that the model can accurately differentiate between positive and negative predictions with minimal misclassification. However, the near-perfect shape of the curves might also mean mild overfitting, as the model could be performing exceptionally well on training data but might generalise less effectively on unseen samples. Further optimisation and evaluation using validation loss were therefore essential to confirm the model's generalisation performance.

The next step in assessing potential overfitting involved examining the loss behaviour during training. For LSTM models, loss is defined using the categorical cross-entropy function, which measures the difference between the predicted probability distribution and the true class labels.

A consistent decline in training loss alongside an increase in validation loss typically indicates overfitting. The trend of the model's training and validation loss is illustrated in Figure 6.17.

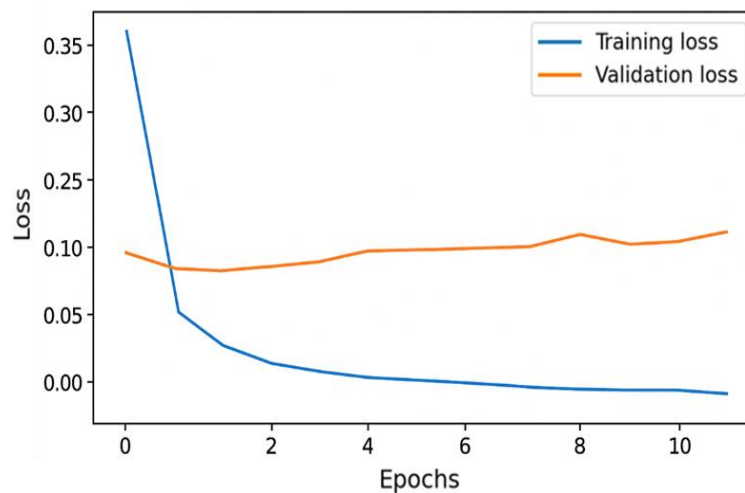


Figure 6.17: Bias-variance trade-off for validation loss of the initial LSTM model

The training results demonstrated that the model was achieving high accuracy and low loss on both the training and validation sets, but there were a few noteworthy points:

- **Accuracy trends:** The training accuracy increased steadily, reaching very high values, indicating that the model was learning effectively during training. The validation accuracy remained strong, peaking at around 97% to 98%. This indicated that the model generalised well to unseen data, but the gap between training and validation accuracy may indicate some degree of overfitting.
- **Loss trends:** The training loss decreased consistently, as expected, as the model learned. However, the validation loss exhibited a different trend. With a few added epochs, the validation loss increased slightly, while the accuracy remained high. This could suggest overfitting, where the model became too specialised to the training data and performed less effectively on unseen data.
- **Overfitting indicators:** An increasing validation loss with stable validation accuracy may indicate that the model was overfitting. Overfitting occurs when the model becomes overly complex and captures noise in the training data rather than the general patterns.

The training loss consistently decreased while the validation loss began to fluctuate and slightly increased with a few more epochs. This confirmed that the model was overfitting, achieving high accuracy on the training data but struggling to generalise to unseen data. The divergence between training and validation performance pointed to a high-variance model. Regularisation techniques could be introduced to reduce overfitting. Alternatively, early stopping and a decreased learning rate could be applied to halt training once the validation loss started to

increase. These adjustments would help improve the model's ability to generalise, achieving a better balance between bias and variance. Therefore, in the revised LSTM model, several key modifications were implemented to optimise performance and improve generalisation capabilities. The bounds of the hyperparameters are listed in Table 6.10.

Table 6.10: Upper and lower bounds of hyperparameters in the revised LSTM model

Hyperparameter	Lower bound	Upper bound	Explanation
Number of LSTM units	8	64	The number of units per LSTM layer was increased from 8 in the initial model to 64 in the revised model. This adjustment enhanced the model's ability to capture complex temporal dependencies and richer feature representations.
Dropout rate - first LSTM layer	0.5	0.5	A fixed dropout rate of 0.5 was applied to the first LSTM layer to provide moderate regularisation by randomly deactivating half of the neurons during training.
Dropout rate - second LSTM layer	0.8	0.8	A higher dropout rate of 0.8 was applied to the second LSTM layer, intensifying regularisation to reduce overfitting and improve generalisation
Learning rate	0.00005	0.00005	The learning rate for the Adam optimiser was fine-tuned to this value to ensure gradual and stable convergence during training.
Early stopping patience	1	3	The early stopping mechanism was modified from a patience of 1 to a patience of 3, allowing the model to continue training for up to three additional epochs after a rise in validation loss before halting.

These adjustments aimed to mitigate overfitting while improving the model's ability to generalise to unseen data. The revised model is visualised in Figure 6.18.

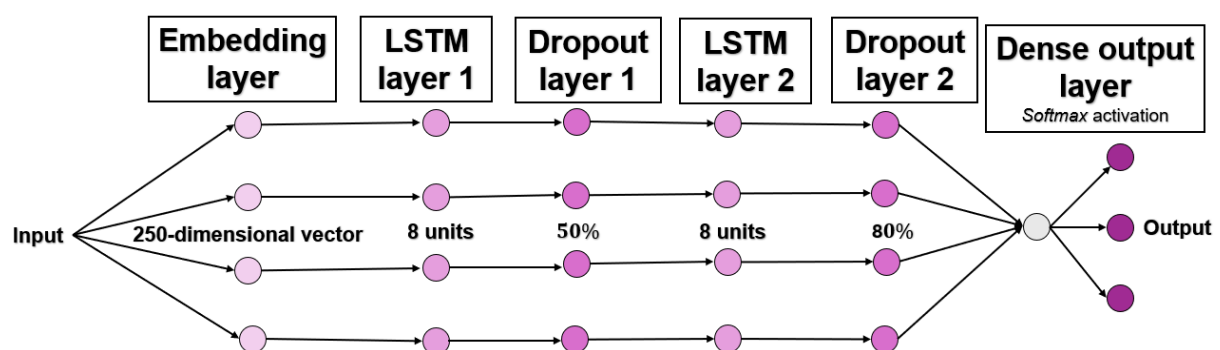


Figure 6.18: Visualisation of the revised LSTM model

The revised LSTM model introduced several structural and optimisation changes to enhance performance and reduce overfitting. The architecture retained the 250-dimensional embedding layer but added a second LSTM block, allowing the model to learn deeper temporal patterns by

stacking two LSTM layers with eight units each. To further improve generalisation, two dropout layers were incorporated: a 50% dropout rate after the first LSTM layer and an 80% dropout rate after the second, each randomly deactivating a proportion of units during training to limit over-reliance on specific neurons. In addition to these architectural refinements, the model was trained using the Adam optimiser with a reduced learning rate of 0.00005 and an early stopping criterion with a patience of three, ensuring stable convergence while preventing unnecessary training cycles. These optimisation adjustments collectively improved model robustness and enabled better generalisation to unseen data. The performance of the revised model is shown in Figure 6.19.

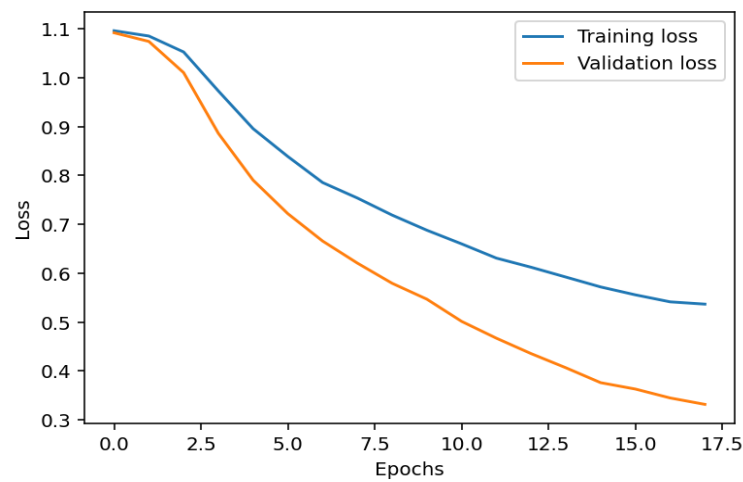


Figure 6.19: The bias-variance trade-off for validation loss of the revised LSTM model

The trends of training and validation loss over several epochs for the revised LSTM model emphasised the balance between the bias and variance. A high bias, which indicates underfitting, is characterised by both training and validation losses that remain high. Conversely, a high variance, which suggests overfitting, is evident when the training loss is significantly lower than the validation loss. In this case, both loss values decreased steadily, reflecting effective learning throughout the training process. Although the model was set to train for 20 epochs, early stopping halted training at epoch 18 because the validation loss plateaued. This indicated that no further improvement was achievable and that continued training posed a risk of overfitting. Early stopping effectively ensured that the model generalised well to unseen data, thus balancing performance while avoiding unnecessary overtraining. Furthermore, the slightly lower validation loss relative to training loss is attributable to dropout regularisation applied during training, which increases training loss while validation is evaluated using the full network structure. The performance of the revised model is presented in Figure 6.20.

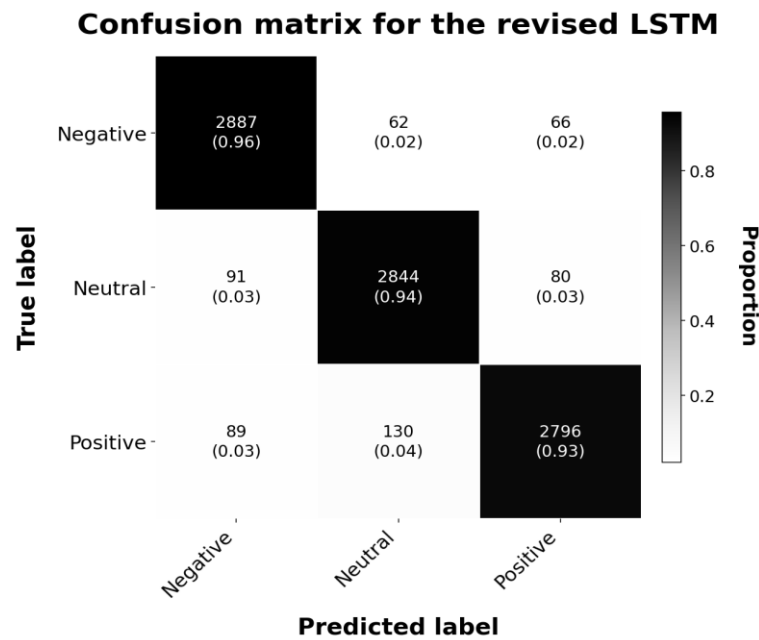


Figure 6.20: Confusion matrix for the revised LSTM model

The confusion matrix for the revised LSTM model demonstrates that the model maintained a strong predictive accuracy across all sentiment classes while showing slight improvements in classification stability compared to the initial model. The majority of predictions were concentrated along the diagonal cells, indicating correct classifications for negative, neutral, and positive sentiments, with proportions of 0.96, 0.94, and 0.93, respectively. This consistency reflected effective generalisation and reduced variance between classes after hyperparameter optimisation.

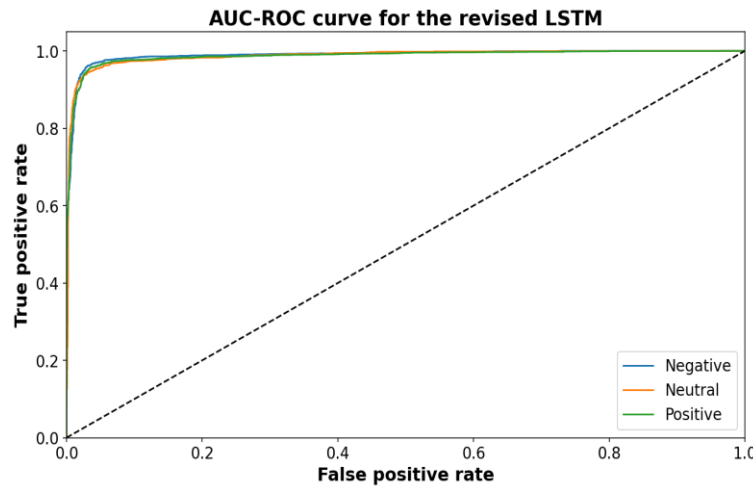
The introduction of regularisation techniques increased embedding units, and an adjusted early stopping patience contributed to more balanced learning, as evidenced by the minimal false positives and false negatives across classes. The improvement in the *negative* class performance, in particular, showed an enhanced sensitivity in distinguishing subtle polarity differences that may have previously caused misclassification. Overall, the revised model achieved a more stable and generalised performance without significant overfitting, which supported the effectiveness of the optimisation adjustments in refining the LSTM architecture. The performance measures for the revised model are summarised in Table 6.11.

Table 6.11: Classification report for the revised LSTM model

Sentiment	Metrics							
	Accuracy	Change	Precision	Change	Recall	Change	F1-score	Change
Negative	0.90	+0.03	0.94	+0.03	0.96	+0.04	0.95	+0.03
Neutral	0.89	+0.02	0.94	−0.02	0.94	−0.02	0.94	−0.02
Positive	0.89	+0.02	0.95	+0.01	0.93	+0.00	0.94	+0.00

The accuracy of all three sentiments increased, with the largest gain observed for the negative class (+0.03). It reflected the model's enhanced ability to identify negative sentiments after optimising correctly. Precision and recall also improved notably for the negative class, which indicated a more balanced trade-off between FPs and FNs. Although the neutral class experienced a slight decline in precision and recall (−0.02), this small reduction suggested that the revised model was more conservative in classifying neutral sentiments, possibly to reduce misclassifications with the polar classes. The positive class maintained a stable recall and showed marginal gains in precision and accuracy, implying consistent predictive performance. The overall accuracy was 0.94.

Overall, the results confirmed that the model adjustments effectively enhanced generalisation, particularly by reducing overfitting and achieving more uniform and reliable sentiment discrimination. The AUC-ROC curve for the revised model is shown in Figure 6.21.

**Figure 6.21: AUC-ROC curve for the revised LSTM model**

The near-perfect area AUC intimated that the revised model effectively distinguished between different sentiment categories with minimal misclassification. The improved smoothness and consistency of the ROC curves compared to the initial LSTM model demonstrated enhanced stability and generalisation after hyperparameter optimisation. This improvement aligned with the reduced validation loss and more consistent classification metrics. It also confirmed that the

regularisation adjustments, such as modified dropout rates, early stopping, and increased output units, successfully mitigated overfitting and strengthened the model's predictive reliability.

All three models discussed in this section demonstrated strong classification performance. However, the LSTM model achieved the most balanced and superior results across all evaluation metrics. Moreover, it achieved an optimal balance between precision and recall. Compared to the logistic regression and random forests models, which exhibited slightly higher biases toward the neutral class and signs of limited generalisation, the final LSTM model's architecture enabled it to capture sequential dependencies and contextual nuances in textual data more effectively. These results confirmed that the LSTM model not only achieved the highest overall predictive consistency but also generalised better to unseen data, making it the most robust model for sentiment classification in this study.

6.4 Final model

Stratified k -fold cross-validation was performed to assess the stability and generalisability of the model by splitting the dataset into multiple training and validation subsets. It is a widely used resampling technique for model evaluation. In this approach, the dataset is generally divided into k equally sized subsets, or folds, to ensure that each observation has the opportunity to appear in both the training and validation sets. This method helped evaluate how well the model performed across different samples of the data, reducing the risk of overfitting to a particular subset.

In this study, $k = 10$ was selected, which represented a well-established balance between computational efficiency and reliability of performance estimation. A 10-fold configuration provided sufficient variability to reduce bias while maintaining manageable variance in the results. Each fold served as the validation set once, while the remaining nine folds were used for training. This iterative process produced ten independent performance estimates, which were then averaged to yield a more stable and generalised measure of model performance. Figure 6.22 illustrates the different splits of the 10-fold cross-validation approach.

Each shaded block represented a fold that was temporarily withheld as the validation set during one learning iteration. Although the illustration shows these validation blocks as distinct and non-overlapping for conceptual clarity, in practice, the randomly-selected validation samples may partially overlap across folds depending on the sampling strategy.

$k = 1$	10%								
$k = 2$		10%							
$k = 3$			10%						
$k = 4$				10%					
$k = 5$					10%				
$k = 6$						10%			
$k = 7$							10%		
$k = 8$								10%	
$k = 9$									10%
$k = 10$									

Figure 6.22: Example of 10-fold cross-validation

The purpose of the diagram is therefore to demonstrate the sequence of fold rotations. Therefore, each subset was independently validated while contributing to the overall evaluation of model robustness and generalisation capability. Figure 6.23 presents the validation accuracy values obtained across the different folds for the LSTM model.

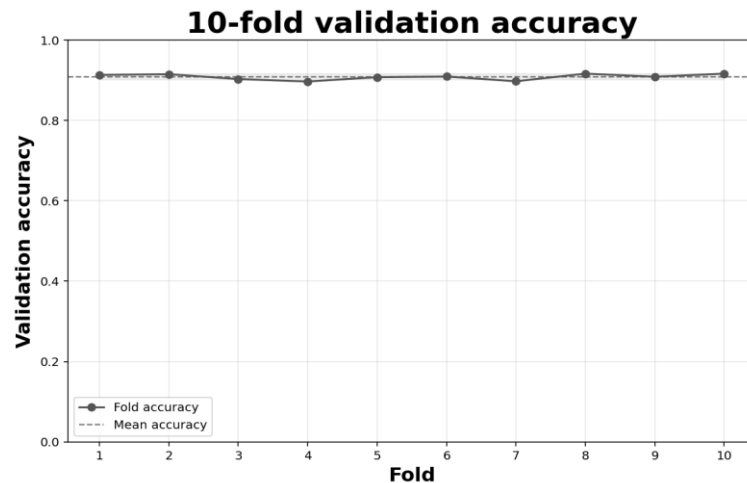


Figure 6.23: 10-fold cross-validation accuracies of the LSTM model

The analysis showed an average accuracy of 93.70%, with a standard deviation of 1.40% and a variance of 0.0196%. These results indicated that the model's performance was consistent across different subsets, demonstrating its robustness and stability. The small variance suggested minimal fluctuations in accuracy, meaning the model generalised well to unseen data.

The revised LSTM model demonstrated the best performance in all aspects of sentiment classification. The AUC-ROC curve highlighted the model's effective ability to discriminate between negative, neutral and positive sentiments with minimal false positives. The classification report confirmed a balanced performance with high precision, recall and F1-scores across all sentiment classes. Additionally, cross-validation results further validated the model's robustness and stability, with consistent performance across different data subsets. Overall, the model exhibited a high level of generalisability and accuracy, making it a reliable tool for multiclass sentiment analysis, catering to a specific trading strategy.

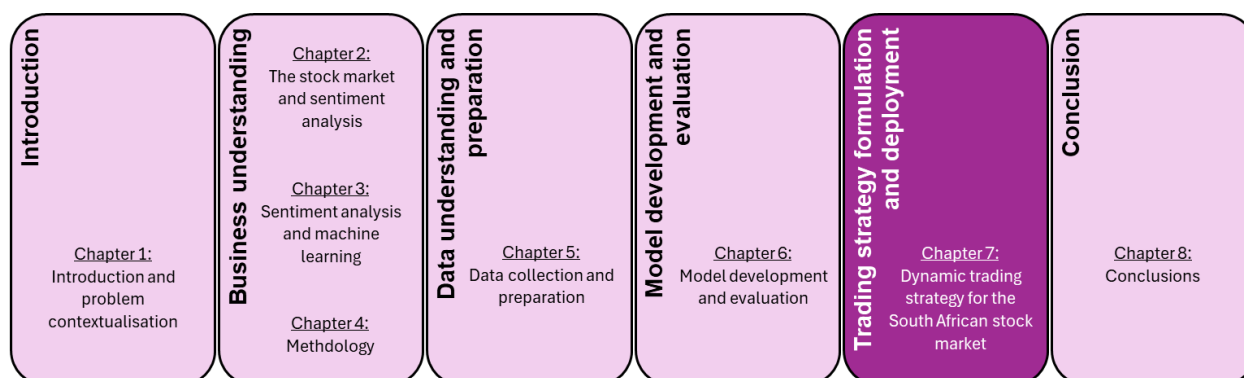
6.5 Conclusion

This chapter provided a detailed evaluation of logistic regression, random forests and the LSTM models for sentiment analysis in the context of stock market predictions. The models were trained and evaluated on social media sentiment data, with particular attention to the balance of sentiment distributions and the impact of pre-processing techniques.

The findings from Chapter 6 provide a solid foundation for applying sentiment analysis to real-world financial contexts. In the next chapter, these results are employed to develop a dynamic trading strategy for the South African stock market. The purpose of the strategy is to translate media-derived sentiment and emotion into actionable trading signals and evaluate their practical performance across selected JSE stocks.

CHAPTER 7

Dynamic trading strategy for the South African stock market



7.1	Introduction	171
7.2	Keyword extraction and strategic mapping.....	171
7.2.1	Extraction of keywords by sentiment category	171
7.2.2	Mapping to strategic frameworks	174
7.3	Stock-specific relationship analysis results.....	178
7.3.1	Hammerson PLC	178
7.3.2	Redefine Properties	179
7.3.3	FirstRand Limited	180
7.3.4	Equites Property Fund	181
7.3.5	Finbond Group Limited	182
7.3.6	Growthpoint Properties	183
7.3.7	Standard Bank Group	184
7.4	Emotional valence and cluster analysis	185
7.4.1	Emotional associations with sentiment	186
7.4.2	K-means clustering of emotional signals.....	189
7.5	Thresholds for sentiment intensity classification.....	194
7.6	Trading signals.....	195
7.7	The dynamic trading strategy.....	196
7.8	Testing and evaluation.....	200
7.8.1	Approach	200
7.8.2	Results	201
7.9	Conclusion.....	202

7.1 Introduction

A dynamic trading strategy for the South African stock market was developed by converting media-derived sentiment and emotion into actionable buy, sell, or hold signals. The process of transforming sentiment-rich textual data into meaningful insights related to stock market trading, is discussed in Section 7.2. The approach involved extracting domain-relevant keywords and situating them within well-known business and strategic analysis frameworks. This is followed by a review of the sentiment relationship analysis in Section 7.3, which focuses on operational practicality and robustness across seven prominent stocks listed on the Johannesburg Stock Exchange (JSE). Section 7.4 presents findings on the relationship between sentiment and market behaviour using correlation, lag, and price movement evidence, and includes the quantification of emotional valence with k-means clustering. Consequently, the calibration of the data-driven thresholds to distinguish strong, weak, and neutral intensities follows in Section 7.5. Section 7.6 covers the machine learning testing phase, in which sentiment and emotion were integrated into a transparent trading rule and evaluated under realistic cost and liquidity assumptions. The development and testing of the dynamic trading strategy is covered in Section 7.7 and 7.8. From these sections, it is evident that the results inform readiness for deployment in a production trading environment.

7.2 Keyword extraction and strategic mapping

Each sentiment was associated with specific words that reflect the underlying emotions expressed in social media discussions. These words served as markers that informed the feelings and thoughts supporting the diverse discussions that occurred online.

7.2.1 Extraction of keywords by sentiment category

Keyword extraction was performed through a word-level sentiment analysis pipeline. After pre-processing, each remaining word was individually scored using the Valence Aware Dictionary for Sentiment Reasoning (VADER) sentiment analyser. The words were classified into one of three categories, namely negative, neutral, or positive. Word frequencies were calculated using a token-level counter, and only words with more than two occurrences were retained to ensure reliability. This methodology allowed for the empirical mapping of keywords to sentiment categories based on both their frequency and polarity strength, providing a data-driven foundation for subsequent sentiment modelling.

To present the results in a scientifically interpretable format, tables are provided for each sentiment category. These tables provide a structured overview of the most representative and sentiment-intense keywords associated with each sentiment class, offering insights into the

emotional language patterns embedded in online news headlines. Table 7.1 presents the words associated with a negative sentiment that had the highest frequency and strongest sentiment score.

Table 7.1: Keywords associated with a negative sentiment

Frequently used words			Strongest negative sentiment		
Word	Frequency	Average sentiment score	Word	Frequency	Average sentiment score
injured	1123	−40.19%	rapist	44	−70.96%
seriously	1005	−17.79%	slavery	4	−70.03%
arrest	804	−34.00%	murder	756	−69.08%
murder	756	−69.08%	kill	654	−69.08%
kill	654	−69.08%	rape	114	−69.08%
warn	580	−10.27%	terrorist	9	−69.08%
pay	579	−10.27%	murderer	27	−68.08%
suspect	474	−29.60%	hell	15	−68.08%
death	471	−59.94%	terrorism	8	−68.08%
fire	468	−34.00%	suicide	46	−67.05%
leave	375	−5.16%	fatality	29	−67.05%
attack	371	−47.67%	suicidal	5	−67.05%
miss	350	−15.31%	killed	4	−67.05%
dead	350	−64.86%	horrific	72	−65.97%
lose	332	−40.19%	cancer	39	−65.97%

The table shows that negative sentiment in news headlines was often centred on themes of violence, crime, and loss, with words like 'injured', 'murder', 'kill', 'attack', and 'death' appearing most frequently. Common terms such as 'arrest', 'suspect', and 'fire' suggested recurring coverage of law enforcement and emergency events, while context-dependent words like 'leave' and 'miss' indicated that negativity sometimes depended on how a story was framed.

The table also lists those words with the strongest negative sentiment scores, such as 'murder', 'rape', 'suicide', and 'terrorist'. Even though some of these appeared less often, they carried much stronger emotional weight, reflecting the intensity of negative language in the news. Overall, both the frequent and emotionally-charged words revealed how strongly media narratives focused on themes of security, violence, and human suffering. Table 7.2 lists the most frequently occurring keywords that received a neutral sentiment classification. They generally represented geographic locations ('south', 'africa', 'cape', 'ballito', 'town'), proper nouns ('eskom', 'ramaphosa'), and context-specific terms such as 'load', 'shed', and 'schedule'.

Table 7.2: Keywords associated with neutral sentiment

Frequently used neutral sentiment keywords					
Word	Frequency	Average sentiment score	Word	Frequency	Average sentiment score
south	6258	0%	ramaphosa	1159	0%
africa	4559	0%	world	1091	0%
african	2199	0%	wave	1087	0%
cape	2000	0%	year	1068	0%
load	1893	0%	schedule	1040	0%
shed	1888	0%	surfer	997	0%
eskom	1515	0%	ballito	996	0%
town	1311	0%			

These words were often associated with neutral reporting on national events or infrastructure. The absence of emotional polarity in these terms reflected their informational or factual nature, rather than any inherent affective content. Notably, the inclusion of terms like 'wave', 'year', and 'world' suggested the presence of temporal or global framing in the headlines, without strong emotional connotations. This aligned with expectations, as neutral words were often descriptive or contextual anchors in news language, rather than expressions of emotion or opinion.

In Table 7.3, the words associated with a positive sentiment are listed. The focus was on the highest frequency and strongest sentiment scores.

Table 7.3: Keywords associated with positive sentiment

Frequently used words			Strongest positive sentiment		
Word	Frequency	Average sentiment score	Word	Frequency	Average sentiment score
grant	672	36.12%	euphoria	3	64.86%
party	483	40.19%	love	138	63.69%
increase	407	31.82%	freedom	56	63.69%
rich	372	55.74%	paradise	10	63.69%
interest	314	45.88%	perfectly	3	63.69%
win	294	58.59%	excellence	6	62.49%
save	244	49.39%	outstanding	9	61.24%
rescue	195	51.06%	super	66	59.94%
security	188	34.00%	champion	49	59.94%
worth	183	22.63%	beautiful	46	59.94%
holiday	178	40.19%	delight	16	59.94%
growth	174	38.18%	brilliance	5	59.94%
support	172	40.19%	joyful	4	59.94%
energy	172	27.32%	congratulation	3	59.94%
challenge	171	7.72%	magnificent	3	59.94%

Many of these reflected themes of improvement, stability, and opportunity, such as 'increase', 'growth', 'support', and 'energy', while others, like 'win', 'rescue', 'save', and rich conveyed success, protection, or prosperity. Words such as 'holiday' and 'party' suggested leisure and celebration, and 'grant' and 'interest' pointed to frequent coverage of economic policy or financial support. Strongly positive terms like 'congratulation', 'champion', 'love', and 'freedom' carried high emotional intensity, highlighting uplifting and aspirational language in stories about achievement, pride, and human resilience.

The analysis resulted in the identification of the most representative keywords across the three sentiment categories based on frequency and sentiment strength. These keywords not only revealed dominant emotional tones embedded within the news headlines but also reflected the public discourse surrounding socio-political, economic, and institutional themes. The exploration laid the foundation for interpreting how these keywords connected to broader strategic forces.

The analysis was extended by mapping the extracted keywords to the PESTEL macro-environmental (Yüksel, 2012) and Porter's Five Forces micro-competitive frameworks (Porter, 2008). Such alignment offered greater comprehension into how emotional language might mirror, respond to, or even shape external market conditions and institutional pressures.

7.2.2 Mapping to strategic frameworks

To identify the underlying drivers of sentiment within the analysed corpus, the extracted keywords were aligned with established business analysis frameworks. This mapping allowed for a structured interpretation of how emotionally charged language reflects broader environmental and competitive dynamics.

7.2.2.1 PESTEL analysis

The keyword extraction process allowed for an opportunity to interpret linguistic patterns specifically related to the PESTEL strategic macro-environmental forces. A selection of high-impact sentiment keywords was evaluated across the following dimensions: Political, Economic, Social, Technological, Environmental, and Legal. This provided a useful framework for classifying and understanding how frequently occurring and emotionally-charged words reflected wider systemic influences present in online news media.

The keyword 'regulation', for example, which was frequently associated with negative sentiment, was mapped to both the Political and Legal dimensions. Its co-occurrence with terms such as 'uncertainty', 'restrictive', and 'compliance burden' signalled apprehension around policy interventions and evolving governance requirements. Similarly, 'inflation', another negatively

charged keyword, clearly reflected economic concerns that included rising costs, reduced purchasing power, and eroded margins.

In contrast, the keyword 'innovation', which carried a strong positive sentiment, aligned with the Technological dimension. It suggested market optimism about improvements in products, services, or operational efficiency. Other keywords, such as 'layoffs' and 'protests', were mapped to the social category, pointing to issues of workforce instability and social unrest. Finally, terms such as 'carbon footprint' and 'sustainability' were clearly aligned with the Environmental dimension and often used in contexts where corporate responsibility and Environmental, Social, and Governance (ESG) concerns were being addressed. Table 7.4 lists the alignment of the identified sentiment-laden keywords with the PESTEL framework.

Table 7.4: Keywords mapped to the PESTEL framework

Keyword	PESTEL analysis		
	Sentiment	PESTEL factor	Explanation
regulation	Negative	Political / Legal	Policy changes or additional compliance requirements are implied, potentially placing pressure on organisational operations.
inflation	Negative	Economic	Rising cost pressures and reduced purchasing power point to broader economic strain.
layoffs	Negative	Social	Workforce reductions indicate instability and often heighten concern among employees and communities.
protests	Negative	Social	Public unrest or dissatisfaction is suggested, which may disrupt social or economic activities.
carbon footprint	Negative	Environmental	Ecological and climate-related concerns are highlighted when this term appears in news discourse.
sustainability	Positive	Environmental	Environmental responsibility and long-term ecological alignment are conveyed through this keyword.
innovation	Positive	Technological	Optimism about technological progress, product development, or improvements in processes is reflected here.
freedom	Positive	Social / Political	The term underscores societal values or national identity and may imply stability within the socio-political environment.
grant	Positive	Economic	Financial support provided by the government or institutions is indicated, often enhancing economic prospects.
eskom	Neutral	Technological	References to South Africa's electricity provider point to infrastructure reliability and energy-related operational concerns.

The keywords for this exploration were identified through frequency analysis and sentiment scoring, where terms with strong polarity values or repeated appearance in sentiment-intense headlines were extracted as the most influential contributors. Each keyword was mapped based on the thematic context in which it typically appeared in the dataset, along with its associated sentiment polarity. This classification provided a structured interpretation of how media language reflected broader macro-environmental concerns, including policy shifts, economic pressures, social dynamics, technological advancements, and sustainability discourse.

7.2.2.2 Porter's Five Forces

An additional mapping to Porter's Five Forces framework was performed to interpret micro-environmental and competitive dynamics. These forces include buyer power, supplier power, threat of substitutes, threat of new entrants, and industry rivalry. Keywords with clear strategic connotations, especially those carrying strong sentiment, provided valuable indicators of the perceived market forces.

Recurrent negative sentiment surrounding 'pricing pressure' or 'affordability' might have reflected stronger buyer-power, where consumers have control over pricing strategies. Similarly, keywords such as 'supply chain disruption' or 'shortage' might have signalled structural vulnerabilities arising from supplier dominance and critical upstream dependencies. Terms like 'alternative platforms' or 'new entrant' - even in neutral or positive contexts - could imply the threat of substitutes or new market entries. Emotionally-charged terms such as 'competition', 'market share', or 'price war' could indicate intense industry rivalry. Table 7.5 presents some of the selected keywords mapped to Porter's Five Forces framework.

Each keyword in the table reflects the underlying market dynamics such as buyer or supplier influence, the threat posed by substitutes or new entrants, and the degree of industry rivalry. These keywords were selected through frequency analysis and sentiment scoring, where terms repeatedly associated with competitive pressure, pricing dynamics, supplier constraints or substitute offerings were extracted as the most influential indicators. This classification enhanced the interpretive value of these keywords by indicating whether the language conveyed perceived risk, opportunity, or neutrality in market conditions. The mapping illustrated how emotional tone in news coverage could reveal stakeholder pressure points and support real-time strategic interpretation of competitive threats.

Table 7.5: Keywords mapped to Porter's Five Forces

Porter's Five Forces analysis			
Keyword	Sentiment	Porter Force	Explanation
pricing pressure	Negative	Buyer Power	Customers appear to exert a strong influence over pricing decisions which place downward pressure on margins.
supply chain disruption	Negative	Supplier Power	Upstream vulnerabilities and delays signal disruptions that may constrain production or service delivery.
alternative platforms	Neutral/Positive	Threat of Substitutes	Innovation from competing solutions introduces viable alternatives that can shift consumer preferences.
new entrant	Neutral	Threat of New Entrants	Emerging competitors entering the market suggest the potential for increased competition and reduced dominance.
competition	Mixed	Industry Rivalry	The keyword reflects varying levels of competitive intensity within the industry, ranging from moderate to heightened rivalry.
price war	Negative	Industry Rivalry	Aggressive pricing behaviour points to firms undercutting one another, which often results in margin compression.
market share	Mixed	Industry Rivalry	Focus on expanding or defending market position demonstrates strategic rivalry among industry players.
pricing pressure	Negative	Buyer Power	Customer leverage over pricing decisions is evident, creating challenges for maintaining profitability.
supply chain disruption	Negative	Supplier Power	Vulnerabilities in supply networks indicate risks that may hinder operational stability.
alternative platforms	Neutral/Positive	Threat of Substitutes	New or improved platforms provide alternative options that can increase competitive pressure.

Sentiment-laden keywords were considered not merely linguistic artefacts but reflected and reinforced both the macro-level environmental trends and micro-level market dynamics. These analyses supported public discourse, perceived risks, and strategic positioning of the identified keywords within the field of stock trading in a specific environment. Sentiment generally not only mirrors external forces but can also actively shape business responses, investor behaviour, and industry evolution in real time.

7.3 Stock-specific relationship analysis results

The analyses were conducted for the training and validation data sets, which spanned the period of January 2018 to December 2023. The testing set includes unseen data from January 2024 to June 2024. It was noted early in the test stages that the COVID-19 pandemic had a notable impact on the results. Therefore, for realistic implementation potential, the analysis was divided into three distinct segments: pre-COVID-19, COVID-19 and post-COVID-19.

Despite the consistently positive correlations within these periods, the observed negative correlation for the entire time could be explained by shifts in sentiment and market dynamics over time. Sentiment and stock price movements exhibited a structured and coherent relationship that resulted in high positive correlations. Conversely, aggregating these sub-periods introduced inconsistencies, such as changes in sentiment trends, variations in rank distribution, and the disruptive influence of external events like the pandemic. These factors changed the continuity of ranks across the entire dataset, which was particularly impactful according to Spearman's rank correlation, as this method is sensitive to shifts in rank alignment over time. Additionally, overlapping lags contributed to the discrepancy. Sentiment might affect stock prices at different intervals in each sub-period. This created further misalignment when the data was analysed as one set. These findings highlighted the importance of contextualising correlation analyses within distinct temporal frameworks. Aggregating data across periods with fundamentally different market conditions risked obscuring underlying trends and relationships, emphasising the need for segmented analysis.

Given that the post-COVID-19 data was the most recent, it was best suited for analysing the relationship between sentiment and stock price movements, as it represented the most accurate and recent trends. The results of the relationship analyses for each selected stock form the remainder of this section.

7.3.1 Hammerson PLC

A strong influence of market sentiment on the stock price movements for Hammerson PLC was observed across various market phases. The summary of the relationship analysis is tabulated in Table 7.6

Table 7.6: Relationship analysis for Hammerson PLC stock

	Relationship			
	Correlation	Lag (months)	Position	Reasoning
Total	−76.89%	6	Opening price movement percentage	A strong negative correlation with a six-month lag suggests that increases in sentiment tended to precede declines in the opening price, indicating a delayed counter-movement in market reactions.
Pre-COVID-19	80.62%	3	Opening price movement percentage	A high positive correlation with a three-month lag reflects a stable period where sentiment shifts were followed by corresponding movements in the opening price within a relatively short time frame.
COVID-19	70.37%	6	Closing price movement	The positive correlation and extended lag imply that market reactions were slower and more uncertain during the pandemic, with sentiment aligning more closely to closing price changes over a longer horizon.
Post-COVID-19	78.54%	2	Closing price movement	The strong positive correlation with a shorter two-month lag indicates faster market adjustment, suggesting restored confidence and quicker price responses to sentiment changes.

Overall, Hammerson PLC demonstrated pronounced sensitivity to sentiment, especially during periods of market turbulence. This highlights how quickly changes in consumer and investor confidence can translate into volatility within the retail and property sector.

7.3.2 Redefine Properties

The analysis for Redefine Properties demonstrated the relationship between market sentiment and stock price movements across different time periods. Table 7.7 presents a summary of the correlation and market sentiment analysis for its stock.

Table 7.7: Relationship analysis for Redefine Properties stock

	Relationship			
	Correlation	Lag (months)	Position	Reasoning
Total	−84.43%	6	Closing price movement percentage	The strong negative correlation with a six-month lag suggests that increases in sentiment were typically followed by delayed downward movements in the closing price, indicating an inverse and slow-moving market response across the full period.
Pre-COVID-19	69.92%	6	Opening price movement percentage	A moderate-to-strong positive correlation with a longer lag indicates that before the pandemic, changes in sentiment aligned with later movements in opening prices, but the market required several months to reflect these shifts.
COVID-19	74.15%	3	Opening price movement	During the pandemic, the positive correlation with a shorter three-month lag shows that market reactions accelerated, with sentiment being incorporated into opening prices more quickly despite heightened uncertainty.
Post-COVID-19	74.36%	2	Closing price movement	The strong positive correlation and short two-month lag point to improved market stability, where sentiment changes translated more rapidly into closing price adjustments in the post-pandemic environment.

The stock for Redefine Properties, as a major player in the real estate investment trust space, was particularly sensitive to broader market perceptions, with sentiment proving to be a crucial but time-lagged driver of stock performance under evolving economic conditions.

7.3.3 FirstRand Limited

Consider Table 7.8. The relationship between market sentiment and FirstRand Limited's stock price movements over the study period is shown.

Table 7.8: Relationship analysis for FirstRand Limited stock

	Relationship			
	Correlation	Lag (months)	Position	Reasoning
Total	−56.90%	5	Opening price movement	The moderate negative correlation with a five-month lag indicates that increases in sentiment were often followed by opposite movements in opening prices after a delayed period, reflecting a gradual inverse response across the full time-frame.
Pre-COVID-19	73.81%	6	Opening price movement percentage	A strong positive correlation with a six-month lag suggests that before the pandemic, sentiment shifts aligned with later opening price movements, although market adjustments occurred slowly.
COVID-19	72.56%	6	Closing price movement	The strong positive correlation and extended lag imply that during the pandemic, sentiment still influenced closing prices, but the reaction time remained slow due to elevated uncertainty and volatility.
Post-COVID-19	70.27%	2	Closing price movement percentage	A strong positive correlation combined with a short two-month lag indicates that post-pandemic market conditions stabilised, allowing sentiment to translate into closing price changes more quickly.

As a leading South African banking institution, FirstRand's stock performance signalled the evolving impact of sentiment under varying economic conditions, with delayed yet significant reactions to market perception, particularly during times of instability and transition.

7.3.4 Equites Property Fund

To extend the examination of sentiment effects within the property sector, Equites Property Fund offers an informative case. The results of its correlation and market sentiment analysis are presented in Table 7.9.

Table 7.9: Relationship analysis for Equites Property Fund stock

	Relationship			
	Correlation	Lag (months)	Position	Reasoning
Total	–72.09%	6	Opening price movement percentage	The strong negative correlation with a six-month lag suggests that increases in sentiment were typically followed by delayed declines in opening prices, indicating an inverse long-term relationship across the full period.
Pre-COVID-19	76.35%	2	Closing price movement percentage	A strong positive correlation with a short two-month lag signals a period of stability where sentiment shifts were reflected in closing prices relatively quickly before the pandemic.
COVID-19	74.88%	4	Opening price movement	The strong positive correlation with a four-month lag suggests that, despite uncertainty during the pandemic, the market still aligned with sentiment, although reactions to opening prices were slower than usual.
Post-COVID-19	58.03%	2	Closing price movement	The moderate positive correlation and short lag suggest that after the pandemic, market responses became quicker but slightly less aligned with sentiment compared to earlier periods.

Equites Property Fund, with its focus on logistics and industrial real estate, remained sensitive to shifts in sentiment, especially during disruptions that altered the outlook for commercial property demand and stability.

7.3.5 Finbond Group Limited

A consistently strong influence of sentiment was found on the stock price movement of Finbond Group Limited throughout the previously defined phases. In Table 7.10, a summary of the correlation and market sentiment analysis is presented.

Table 7.10: Relationship analysis for Finbond Group Limited stock

	Relationship			
	Correlation	Lag (months)	Position	Reasoning
Total	−68.61%	6	Opening price movement percentage	The strong negative correlation with a six-month lag indicates that increases in sentiment were generally followed by delayed declines in opening prices, suggesting an inverse and slow-reacting relationship over the entire study period.
Pre-COVID-19	72.12%	6	Opening price movement percentage	A strong positive correlation with a long six-month lag reflects a period where sentiment aligned with later opening price movements, but market adjustments occurred gradually before the pandemic.
COVID-19	71.17%	6	Closing price movement	The positive correlation with the same extended lag suggests that during the pandemic, sentiment continued to influence closing prices, although heightened uncertainty slowed the market's reaction time.
Post-COVID-19	80.45%	6	Closing price movement	The very strong positive correlation and long lag imply that sentiment remained highly influential after the pandemic, but price adjustments still occurred with a substantial delay, pointing to slower market normalisation for this stock.

As a specialised lender in the financial services industry, Finbond Group Limited's stock behaviour reflected the delayed but powerful influence of sentiment, particularly during times of uncertainty and eventual recovery.

7.3.6 Growthpoint Properties

Sentiment analysis for the Growthpoint Properties stock demonstrates a strong and sustained relationship between market sentiment and stock price movements across different phases. The findings are outlined in Table 7.11.

Table 7.11: Relationship analysis for Growthpoint Properties stock

	Relationship			
	Correlation	Lag (months)	Position	Reasoning
Total	-76.40%	4	Opening price movement percentage	The strong negative correlation with a four-month lag suggests that increases in sentiment were followed by opposite movements in opening prices, indicating a delayed inverse relationship across the full period.
Pre-COVID-19	81.61%	6	Opening price movement percentage	A very strong positive correlation with an extended six-month lag reflects a stable pre-pandemic environment where sentiment shifts aligned with later opening price movements, although the reaction time was slow.
COVID-19	78.60%	6	Opening price movement	The strong positive correlation and long lag indicate that during the pandemic, sentiment continued to influence opening prices but only after a considerable delay, reflecting heightened uncertainty.
Post-COVID-19	76.02%	2	Closing price movement	The strong positive correlation, combined with a short two-month lag, suggests that in the post-pandemic period, market reactions were faster, and sentiment translated more quickly into closing price adjustments.

As a leading real estate investment trust in South Africa, Growthpoint Properties' stock performance was closely tied to market sentiment, particularly during shifts in economic confidence and sector stability with the data revealing both the strength and delayed nature of this relationship.

7.3.7 Standard Bank Group

The analysis for Standard Bank Group, revealed a strong connection between market sentiment and stock price movements throughout the study period. A summary of the analysis results is displayed in Table 7.12.

As one of South Africa's largest and most systemically important financial institutions, Standard Bank Group's stock performance clearly mirrored the evolving impact of sentiment resulting from online news, marked by strong correlations and delayed yet significant market adjustments during times of both crisis and recovery.

Table 7.12: Relationship analysis for Standard Bank Group stock

	Relationship			
	Correlation	Lag (months)	Position	Reasoning
Total	−63.87%	5	Opening price movement percentage	The moderate negative correlation with a five-month lag indicates that increases in sentiment were often followed by opposite movements in opening prices after a delayed period, reflecting an inverse relationship over the full timeframe.
Pre-COVID-19	75.03%	6	Opening price movement	A strong positive correlation with a long six-month lag suggests that before the pandemic, sentiment aligned with later opening price movements, although market adjustments remained slow.
COVID-19	80.21%	6	Closing price movement	The strong positive correlation and extended lag imply that during the pandemic, sentiment had a significant influence on closing prices, albeit with a delayed effect due to elevated uncertainty and volatility.
Post-COVID-19	69.09%	2	Closing price movement percentage	The strong positive correlation with a short two-month lag indicates quicker market reactions in the post-pandemic period, showing that sentiment translated into closing price movements more rapidly as stability returned.

The results of these analyses showed how sentiment correlated with stock price movements and highlighted the best lags that influenced trading decisions, verifying the model output. These insights were crucial for the next stage in the development of the trading strategy, where the focus shifted to incorporating both sentiment and emotion analysis data to formulate a comprehensive and actionable trading strategy for the selected stocks.

The variation in optimal lags across market regimes likely reflects structural differences in volatility, liquidity, and investor behaviour. Crisis periods such as COVID-19 may accelerate information assimilation due to heightened uncertainty, whereas stable periods may exhibit slower sentiment diffusion. The observed lag shifts, therefore, align with regime-dependent market dynamics rather than purely stochastic variation.

7.4 Emotional valence and cluster analysis

Emotions caused by reading online news significantly influenced market sentiment, reflecting the collective psychological state of investors. Examining variance and average sentiment over different periods provided a clearer picture of how emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust shaped market movements. Each keyword was given a valence score using sentiment lexicons, offering a detailed view of sentiment intensity. K-means

clustering was subsequently applied to group keywords based on their valence scores. This approach helped identify distinct emotional clusters, revealing themes, risk indicators, or trends of optimism in the dataset.

7.4.1 Emotional associations with sentiment

The emotion analysis that was conducted using the *NRC*Lex lexicon, resulted in a mapping of words to eight fundamental emotions that include anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. After undergoing pre-processing, each headline in the dataset was assessed according to its emotional composition. This was followed by a calculation of the frequency of emotion-associated words within each headline, which was expressed as a proportion of the total word count. This resulted in normalised emotion scores for each category. An excerpt, as an example of the emotion analysis output, follows in Table 7.13.

Table 7.13: Summary of characteristics of emotion analysis

Title	Sample output of emotion analysis							
	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
kodak launch cryptocurrency	0	0.5	0	0	0	0	0	0
download game nintendo switch	0	0	0	0	0	0	0	0
way strengthen organisation cloud archive	0	0	0	0	0	0	0	0
bot talk like people	0	0	0	0	0	0	0	0
telegram zero day vulnerability use cryptocurrency mining attack	0.17	0	0	0.33	0	0.17	0	0
european authority investigate austrian bitcoin scam	0	0	0	0	0	0	0	0.33
redstor launch seamless automate cloud backup archive	0	0.5	0	0	0	0	0	0
see btc global scam mile away	0	0	0	0	0	0	0	0
sabc investigate claim guptas acquire archive footage	0	0	0	0	0	0	0	0

sabc coo meeting union unhappy plan commercialise safm	0.2	0.2	0.2	0	0	0.2	0	0
--	-----	-----	-----	---	---	-----	---	---

Each value in the table represents the proportional intensity of an emotion within a given headline, based on the occurrence of emotion-laden words as defined by the NRC lexicon. The granularity of word-level emotion attribution and the varied emotional tone that can be present, even in short-form news content, are visible in the table.

These proportional emotion scores were compiled into a structured dataset that enabled a quantitative analysis of the emotional tone across sentiment categories. Descriptive statistics such as minimum, maximum, average, standard deviation, and variance were computed for each emotion in each sentiment class and are shown in Table 7.14. Insights into the consistency and variability of emotional expressions associated with each sentiment type could be derived.

Table 7.14: Descriptive statistics of emotions across sentiment classes

Emotion	Sentiment	Minimum	Maximum	Average	Standard deviation	Variance
Anger	Negative	0.02643	0.06891	0.05264	0.01117	0.00012
	Neutral	0.05139	0.06957	0.05583	0.00783	0.00006
	Positive	0.03383	0.06854	0.05503	0.00844	0.00007
Anticipation	Negative	0.23103	0.29041	0.25717	0.01657	0.00027
	Neutral	0.23409	0.26319	0.24120	0.01253	0.00016
	Positive	0.22193	0.34429	0.26928	0.03298	0.00109
Disgust	Negative	0.00469	0.03149	0.02108	0.00851	0.00007
	Neutral	0.01667	0.03234	0.02851	0.00675	0.00005
	Positive	0.01131	0.03864	0.02328	0.00688	0.00005
Fear	Negative	0.08369	0.12651	0.10636	0.01182	0.00014
	Neutral	0.10239	0.13556	0.12745	0.01429	0.00020
	Positive	0.05971	0.11408	0.09322	0.01654	0.00027
Joy	Negative	0.02937	0.07070	0.04324	0.01254	0.00016
	Neutral	0.04125	0.04960	0.04329	0.00359	0.00001
	Positive	0.02992	0.05988	0.04465	0.00907	0.00008
Sadness	Negative	0.03272	0.06959	0.05485	0.01000	0.00010
	Neutral	0.04818	0.05069	0.05008	0.00108	0.00000
	Positive	0.03729	0.09000	0.05729	0.01427	0.00020
Surprise	Negative	0.02416	0.06936	0.04052	0.01140	0.00013
	Neutral	0.01987	0.05223	0.04432	0.01394	0.00019
	Positive	0.00632	0.05084	0.03579	0.00810	0.00007
Trust	Negative	0.08454	0.14456	0.10563	0.01560	0.00024
	Neutral	0.08740	0.15625	0.10423	0.02964	0.00088
	Positive	0.07500	0.15876	0.10743	0.01922	0.00037

The values for the *anger*-emotion were similar across sentiment categories, with neutral and positive sentiment displaying slightly higher averages at 0.05583 and 0.05503, respectively, than the negative sentiment at 0.05264. Variability was more pronounced in the negative sentiment at 0.00012. This indicates a lower consistency for a negative sentiment when compared to neutral or positive. *Anticipation* was strongly associated with a positive sentiment at the highest average of 0.26928 and the highest maximum at 0.34429. The negative sentiment followed with an average of 0.25717 while the neutral sentiment showed the lowest anticipation levels at 0.24120. Positive sentiment exhibited the widest range of anticipation, as reflected by its higher variability of 0.00109. This indicated that it was a key emotion during optimistic market phases. The neutral sentiment aligned most strongly with *disgust*, by showing the highest average of 0.02851 and the lowest variability at 0.00005. Negative and positive sentiments had similar average levels of 0.02108 and 0.02328, respectively, with the negative sentiment being slightly more consistent. *Fear* was primarily linked to neutral sentiment, displaying the highest mean value of 0.12745. Negative sentiment followed with an average of 0.10636, while positive sentiment displayed the lowest average at 0.09322, and the highest variability of 0.00027. This meant that it was harder to predict fear in a positive sentiment context. A positive sentiment was clearly tied to *joy*, which showed the highest average at 0.04465 and a maximum of 0.05988. Neutral and negative sentiments exhibited nearly identical averages of 0.04329 and 0.04324, but the negative sentiment had the highest variability at 0.00016. This indicated a greater inconsistency in joy during negative market periods. For the *sadness* emotion, the positive sentiment had the highest average at 0.05729, followed closely by a negative sentiment at 0.05485. Neutral sentiment showed the lowest average of 0.05008 and the least variation at 0, indicating a narrower range of emotional expression. *Surprise* was most strongly associated with a neutral sentiment, with an average of 0.04432 and the variability of 0.00019. Positive sentiment followed with an average of 0.03579, while negative sentiment averaged at 0.04052, but appeared more consistent. The highest average for *trust* was for positive sentiment at 0.10743, followed by negative sentiment at 0.10563. The neutral sentiment confirmed a comparable average of 0.10423 but demonstrated the highest variability at 0.00088. This suggested a broader range of trust-related responses in neutral market contexts.

From the emotional analysis, clear associations emerged:

- Negative sentiment was predominantly linked to fear, sadness, and anger as these emotions consistently exhibited higher averages and lower variability under negative conditions.
- The positive sentiment was closely tied to anticipation and joy as these emotions showed the highest averages and variability. They reflected heightened optimism and confidence during favourable market periods.

- A neutral sentiment was characterised by disgust, surprise, and trust, with peak averages or broader ranges within this category.

The emotion–sentiment relationships provided a nuanced foundation for the clustering procedures in the subsequent phase of this study. By identifying how emotional profiles differed in both intensity and consistency across sentiment classes, the groundwork was laid for constructing more refined sentiment-emotion clusters that might enhance the predictive power of sentiment-based trading strategies.

7.4.2 K-means clustering of emotional signals

K-means clustering is a widely used unsupervised learning method that groups data into a number (k) of clusters based on similarity. The technique works by minimising the within-cluster sum of squared distances between data points and their corresponding centroids, producing compact and well-separated clusters. It is generally an iterative process where centroids are initially assigned, data points are grouped according to their proximity to these centroids, and centroids are then recalculated until stable clusters are formed. The elbow method is a visual technique used to identify the optimal number of clusters in a k-means analysis.

Before applying the algorithm, the data had to be standardised to ensure that variables would contribute equally to the clustering process. This involved plotting the within-cluster sum of squared errors (SSE) against different values of k (in this study, $k \in \{1, \dots, 10\}$) and selecting the point where the rate of decrease in the SSE sharply slowed down, forming an 'elbow' in the curve. The elbow graph for the negative sentiment is visualised/displayed in Figure 7.1.

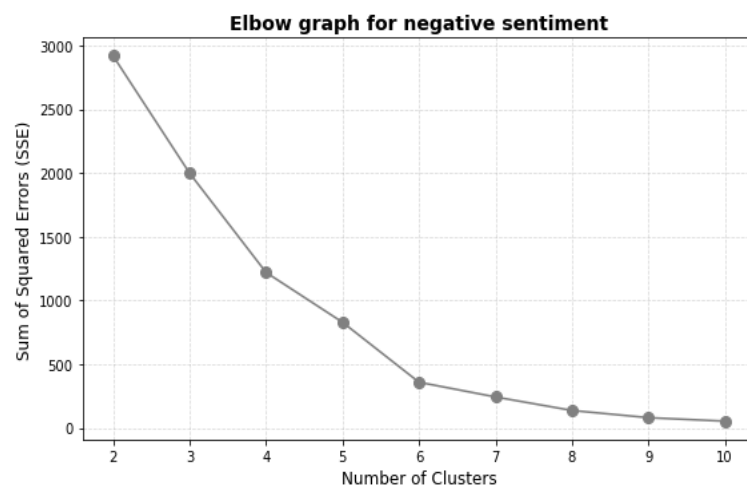


Figure 7.1: Elbow graph for the negative sentiment

In the elbow graph, the SSE sharply decreased as the number of clusters increased, but the rate of decline slowed noticeably after six clusters. The elbow, therefore, formed after six clusters,

which might represent a good balance between minimising the SSE and avoiding overfitting. It was the most suitable choice for this dataset as it balanced interpretability and clustering quality while avoiding unnecessary complexity. The elbow graph for the neutral sentiment is visualised in Figure 7.2.

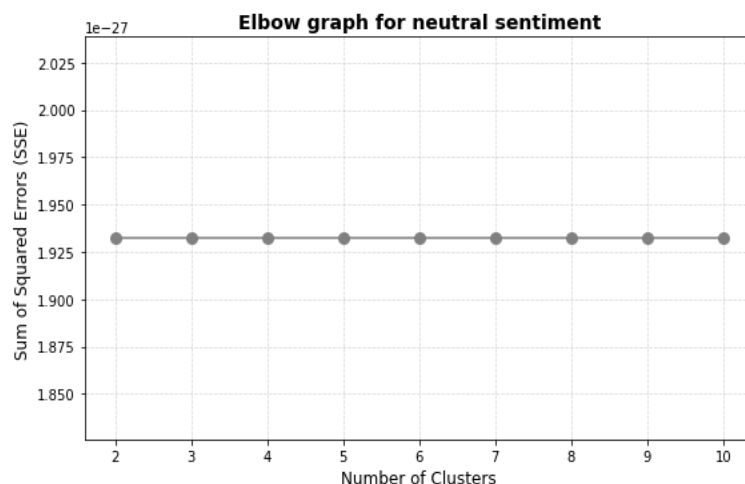


Figure 7.2: Elbow graph for the neutral sentiment

No significant metric variations could be observed across the cluster. The SSE remained constant, and there was no meaningful reduction in intra-cluster variance as the number of clusters increased. This outcome meant that the dataset might inherently lack a meaningful clustering structure or might already have existed as a single cohesive cluster. The elbow graph for the positive sentiment is shown in Figure 7.3.

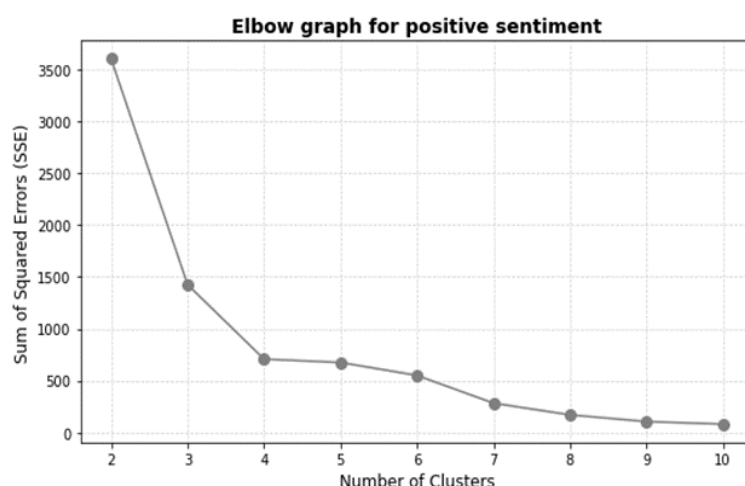


Figure 7.3: Elbow graph for the positive sentiment

A noticeable elbow occurred after about four clusters, where the SSE exhibited a significant decrease initially but started to flatten beyond this point. It was resolved that adding more clusters after this would provide diminishing returns in terms of reducing intra-cluster variance. Therefore,

four clusters were considered the best choice as they balanced a good reduction in SSE and avoided the risk of overfitting while maintaining interpretability.

It was deduced that the optimal number of clusters was six, indeterminate, and four for the negative, neutral, and positive sentiments, respectively. It was noted that the neutral sentiment represented the absence of both positive and negative sentiments. It could be assumed that if the predicted sentiment had been neither positive nor negative and its thresholds were not met, the sentiment was effectively classified as neutral. Once the optimal number of clusters for each sentiment's associated emotions had been identified using the elbow method, *k*-means clustering was applied to group the data points into distinct clusters.

For this analysis, the emotions associated with each sentiment type - negative and positive - were clustered independently to capture the inherent patterns in their distributions. Each cluster represented a grouping of data points with similar emotional intensities that enabled the identification of consistent behavioural trends within each sentiment category. The clustering results demonstrated how specific emotions varied in intensity and under different conditions. For example, clusters for negative emotions resulted in distinct ranges of fear, anger and sadness associated with heightened pessimism in the market. Similarly, clusters for positive emotions revealed patterns of optimism characterised by varying levels of joy and anticipation. By visualising the clusters, their separation and distinctiveness become evident, further confirming the validity of the clustering procedure. Figure 7.4 depicts the clusters of negative sentiment related to anger, fear, and sadness.

The data points are colour-coded according to the assigned clusters, and the scatter plots on the diagonal show the distribution of each sentiment individually. The other plots provide some understanding of the relationships and co-occurrence of the sentiment pairs, revealing patterns in how the negative emotions interact with one another. It was found that different groups of data points aligned with specific emotional characteristics. For example, cluster 0 predominantly comprised observations with higher values of fear. Cluster 5 appeared to have been closely associated with sadness. The remaining clusters displayed varying combinations of fear, anger, and sadness, reflecting different emotional profiles across the dataset.

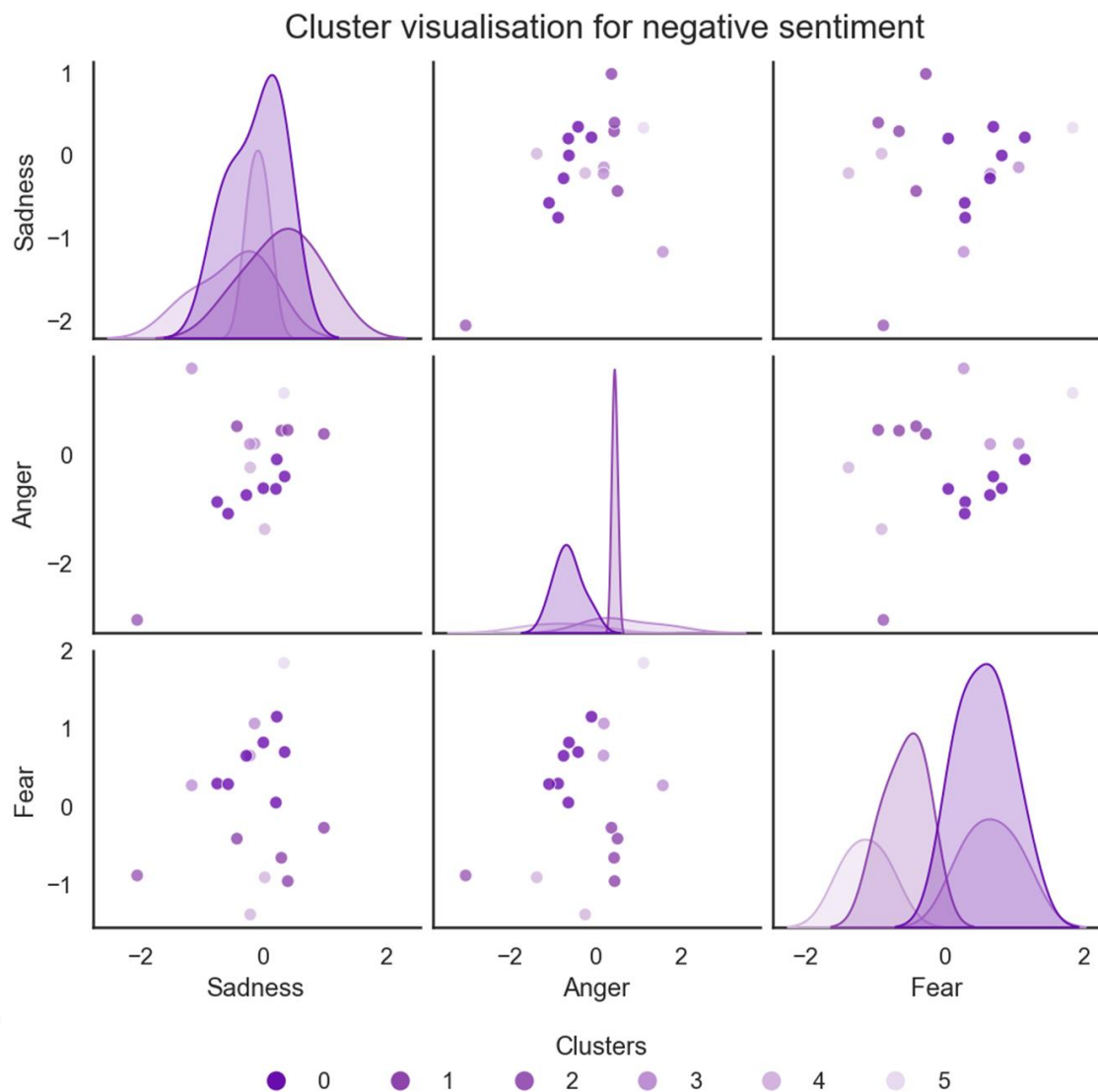


Figure 7.4: The clusters of negative sentiment

Notably, the clustering results revealed that fear and anger tended to form more compact and coherent clusters, implying a closer association with the underlying data. In contrast, sadness displayed a broader dispersion across several clusters, suggesting that its patterns of expression exhibited less clearly defined boundaries compared with the other two emotions. These findings were important for identifying patterns in sentiment that could guide further analysis and interpretation of sentiment data in this context.

Recall that joy and anticipation were identified as the two emotions underlying positive sentiment. Consider the clusters for positive sentiment in Figure 7.5.

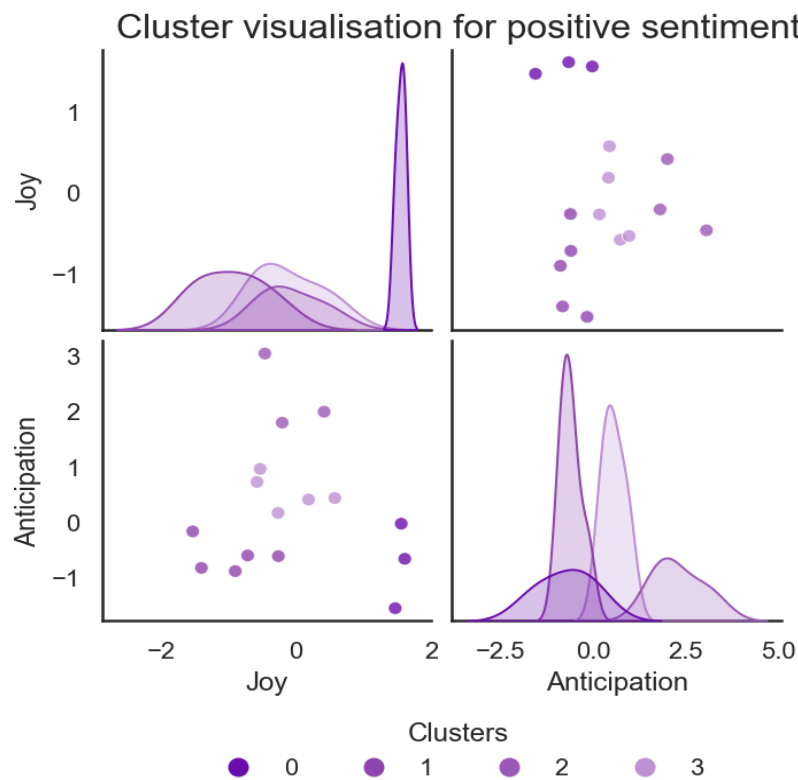


Figure 7.5: The clusters of positive sentiment

Once again, colour-coding was used to differentiate the clusters, with each colour representing a distinct emotional profile. The diagonal subplots show the distribution of each sentiment dimension individually, while the off-diagonal plots reveal the relationships between joy and anticipation. Note the distinct groupings, each characterised by specific emotional characteristics. For example, cluster 0 primarily consists of data points with high values of joy, while cluster 3 shows a concentration of data points with higher anticipation values. Clusters 1 and 2 display mixed expressions of joy and anticipation with different concentration levels along each dimension. A more discrete distribution was observed for joy, with values clustering around extreme positive and negative values. In contrast, for anticipation, a wider spread was visible, especially within clusters 1 and 2, suggesting it may have had a more varied expression in the data. Although their expression patterns differed, joy and anticipation were related.

The identified emotion clusters served as the foundation for determining the thresholds used in trading signal generation. In this context, thresholds were defined as statistically derived boundary values for emotional intensity scores that separate normal emotional variation from significant sentiment shifts. These boundaries, typically calculated using the mean and standard deviation within each cluster, delineated the upper and lower limits for specific sentiment-emotion combinations that drive trading decisions. Consequently, this approach enabled the translation of emotional sentiment into actionable trading signals by identifying when an emotion exceeded or

fell below certain levels. As a result, a dynamic response to sentiment changes in news and corresponding market behaviour could be built into the trading strategy, enabling context-sensitive buy, hold, or sell signal formulation.

7.5 Thresholds for sentiment intensity classification

Development of a dynamic trading strategy involved establishing emotion-specific thresholds that delineated sentiment levels. This section outlines the approach used to derive these thresholds using a combination of clustering and statistical analysis on standardised¹³ emotional variables. After clustering, standardisation involved calculating the statistical properties of each emotional variable within the clusters, including the mean, minimum, maximum, and standard deviation. They formed the basis for quantifying the variability and typical ranges of emotional values within each cluster. In standardised data, the standard deviation equals one across all emotional variables, allowing for meaningful comparison across variables and clusters.

Thresholds were derived using a statistical boundary-based method, commonly referred to as a standard deviation margining approach¹⁴. Grounded in statistical principles, this heuristic value was utilised to exclude central values to emphasise the emotional extremes, where actionable sentiment signals were most likely to occur. Denoting T_{low} and T_{up} as the lower and upper thresholds, respectively, x_{min} and x_{max} as the minimum and maximum asset price movements, respectively, and σ the standard deviation, the following formulas were applied for each emotional variable:

$$T_{low} = x_{min} + \sigma, \text{ and} \quad (7.1)$$

$$T_{up} = x_{max} - \sigma. \quad (7.2)$$

Using this approach, the outer boundaries of variability within each emotion's distribution were captured, identifying high-intensity emotional regions that were more likely to influence market behaviour; for instance, extreme levels of fear or sadness in the negative sentiment cluster were interpreted as stronger emotional signals and therefore used to define buy advice in the trading strategy. The calculated thresholds for both sentiment clusters are shown in Table 7.15.

¹³ Standardised variables are variables that have been transformed to a common scale so that their values are comparable, regardless of the original units or range

¹⁴ The standard deviation margining approach calculates the required margin by adding a risk buffer based on the variability (standard deviation) of asset price movements over a specified period

Table 7.15: Calculated characteristics of negative and positive sentiments

Negative sentiment:					
Emotion	Minimum	Maximum	Standard deviation	Lower threshold	Upper threshold
Anger	−2.35	1.46	1.00	−1.35	0.46
Fear	−1.92	1.70	1.00	−0.92	0.70
Sadness	−2.21	1.47	1.00	−1.21	0.47
Positive sentiment:					
Anticipation	−1.44	2.27	1.00	−0.44	1.27
Joy	−1.63	1.68	1.00	−0.63	0.68

The standard deviation for each emotion consistently equalled 1.00, reflecting the variability within the cluster after standardisation. The thresholds provided a focused range for each emotional variable, delineating the values that were most relevant to the negative sentiment cluster and excluding extreme outliers. Through the implementation of these ranges, precision was ensured, and the effective identification of sentiment-driven trading opportunities was supported.

7.6 Trading signals

Converting sentiment analysis insights into trading signals necessitated a systematic design and validation of signal generation methods. This section details the procedure for developing trading signals based on the quantified emotional valence and sentiment intensity thresholds previously established. The trading rule adopts a contrarian interpretation of extreme sentiment signals, grounded in behavioural finance theory which suggests that markets may temporarily overreact to news. Under this assumption, positive sentiment may indicate short-term overvaluation, while negative sentiment may reflect temporary undervaluation. The strategy, therefore, also tests mean-reversion in addition to momentum effects.

The generation of trading signals relied on the thresholds and the dynamics of short-term and long-term sentiment variability. These signals guided specific actions based on market sentiment:

- Sell signals were associated with positive predicted sentiment clusters where emotions like trust, joy, and anticipation fell within their defined thresholds.
- Buy signals were linked to negative predicted sentiment clusters where emotions such as fear and sadness aligned with their calculated ranges.
- Hold signals were considered when the predicted sentiment was neutral and neither the buy nor sell signal criteria were met.

The integration of sentiment variability, emotional thresholds, and statistical analysis ensured that trading signals provided clear and actionable guidance. These signals enabled an effective response to market sentiment and captured opportunities based on emotional drivers Figure 7.6.

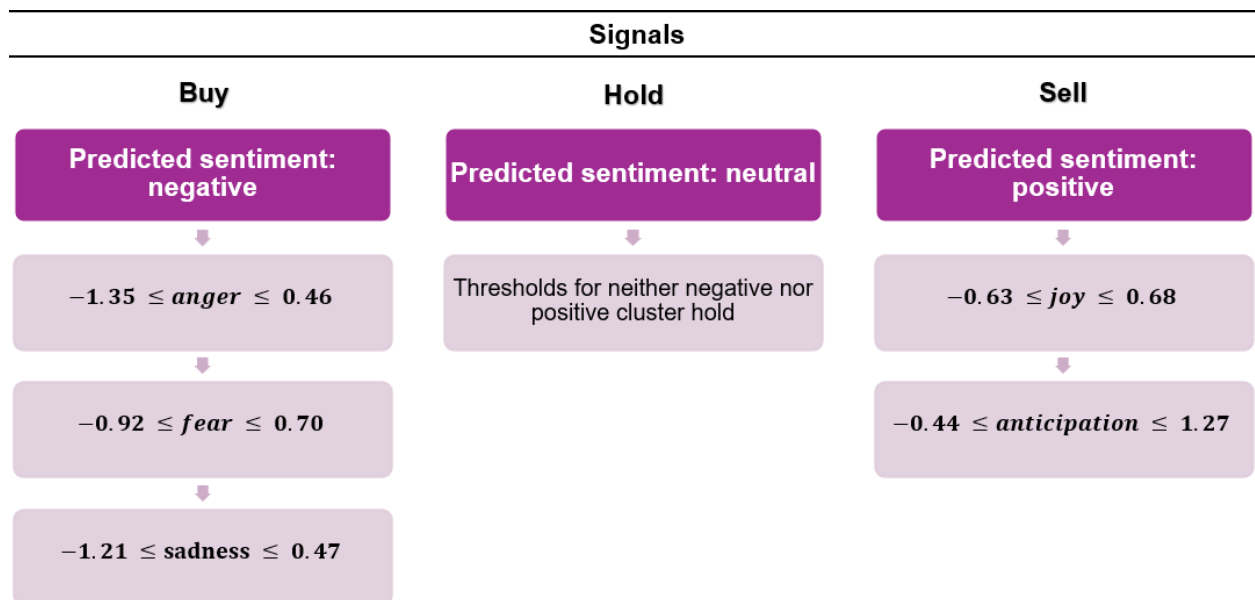


Figure 7.6: The final conditions for trading signal generation

The derived emotional thresholds were applied to guide trading decisions based on sentiment. The thresholds listed in the first column were used to generate a ‘buy’ signal for words identified as predictors of anger, fear, and sadness. Likewise, predictors of joy and anticipation would generate ‘sell’ signals, according to the thresholds in the right column. In cases of neutral sentiment, where the emotional values did not satisfy the thresholds associated with either the negative or positive sentiment clusters, a ‘hold’ signal would be recommended, implying insufficient emotional conviction to warrant immediate action. It is important to note that when none of the sentiment-emotion rules are met, the outcome is a ‘hold’ signal.

Thus, the emotional variables in this figure correspond to those evaluated and used for threshold calculations. Using the final conditions, a dynamic trading strategy could be established for the South African stock market.

7.7 The dynamic trading strategy

Following the series of tests, evaluations, and refinements conducted throughout this study, a dynamic trading strategy was developed that integrates all modelling components into a coherent start-to-end process. The strategy involved the collection of South African news data from selected online platforms, after which the text would undergo comprehensive pre-processing, including lowercasing, noise removal, tokenisation, part-of-speech tagging, and lemmatisation to ensure clean and standardised inputs. Sentiment classification would then be performed using the optimised Long Short-Term Memory (LSTM) model, which would assign a sentiment label to each headline. Building on this, emotion analysis would be applied to determine the emotional signals within each sentiment category, based on the refined emotion-mapping procedure. The

resulting sentiment-emotion pair is to be evaluated through a threshold-based if-then rule that would incorporate the optimised emotional thresholds established previously. Depending on whether these thresholds were exceeded, the strategy would generate a 'buy', 'sell', or 'hold' signal. The complete workflow is summarised in the Unified Modelling Language (UML) diagram shown in Figure 7.7.

Utilisation of the strategy would involve the collection of news headline data, followed by a series of pre-processing steps to ensure the text is clean, structured, and suitable for analysis. Sentiment analysis and emotion analysis would then be conducted in parallel. The sentiment model would assign each headline to one of three categories - negative, positive, or neutral - while the emotion model would quantify the intensity of specific emotions, including anger, fear, sadness, anticipation, and joy.

Trading decisions would be generated by applying threshold-based rules derived from a combination of clustering and statistical analysis. When the predicted sentiment is negative, a 'buy' signal would be issued only if the emotion scores for anger, fear, and sadness simultaneously fall within their respective threshold ranges associated with the negative-sentiment cluster. Conversely, when sentiment is positive, a "sell" signal would be triggered if both anticipation and joy fall within the predefined bounds corresponding to the positive-sentiment cluster. If neither the negative nor the positive threshold conditions were met, a "hold" decision would be produced.

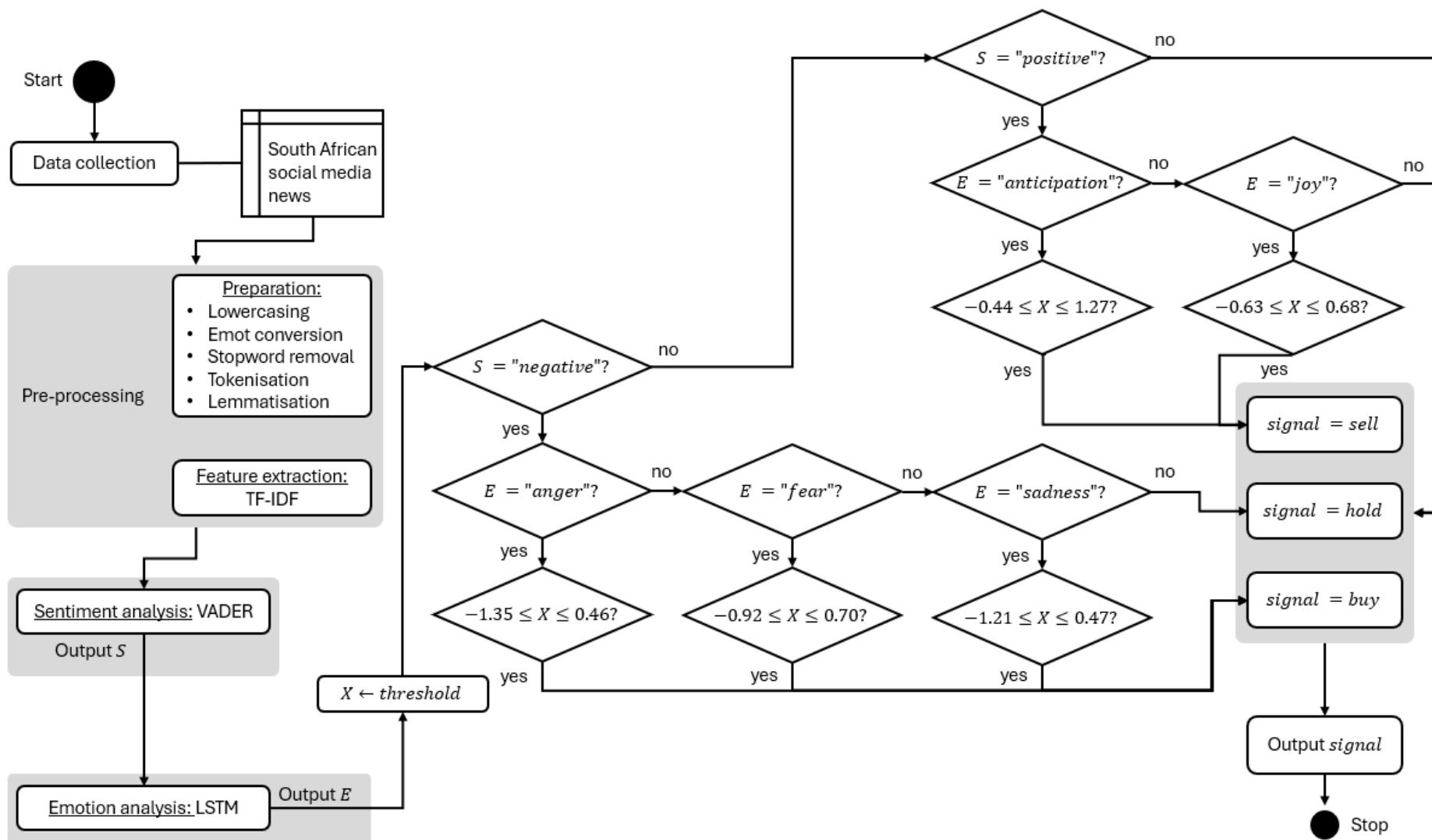


Figure 7.7: The dynamic sentiment-directed machine learning trading strategy

This rule-based strategy would ensure that trading actions were only taken when strong and emotionally coherent signals were present, thereby reducing the likelihood of false or noise-driven trades. The Python pseudocode for this UML diagram follows in Algorithm 7.1.

Algorithm 7.1: Dynamic trading strategy for the South African stock market

Input: *report_month* is the month for which the pipeline is executed;
news_corpus are articles with {date, source, title, body};
price_panel are the month-end stock prices of the seven selected JSE stocks;
preprocess_params include tokenisation, lemmatisation, and stop words;
sentiment_model is the developed sentiment prediction classifier;
emotion_thresholds are the calibrated bounds per emotion;
trade_rules: the range over which to collect historical feed data.

Output: *signals_log* produce sentiment- and emotion-driven buy, sell, or hold signals.

```

1  ##Data collection
2   $A \leftarrow \text{select articles where } publish\_date \in \text{the } report\_month;$ 
3   $P \leftarrow \text{extract month -}$ 
    $\text{end prices for the seven selected stocks from } price\_panel;$ 
4   $T \leftarrow \text{lean } A \text{ using } preprocess\_params;$ 
5  ##Pre-processing
6   $F\_txt \leftarrow \text{construct text features from } T \text{ such as } TF - IDF;$ 
7   $Agg\_txt \leftarrow \text{aggregate features to stock level over the lookback\_window};$ 
8  ##Sentiment analysis and prediction
6   $S\_art \leftarrow \text{apply sentiment\_model to each article to obtain sentiment labels};$ 
7   $S \leftarrow \text{aggregate } S\_art \text{ to a monthly sentiment};$ 
8  ##Emotion analysis
9   $E\_art \leftarrow \text{compute emotion intensity vectors per article};$ 
10  $E \leftarrow \text{aggregate } E\_art \text{ to monthly emotion scores};$ 
11 ##Trading signal
12  $X \leftarrow \text{join } \{Agg\_txt, S, E\} \text{ with } P \text{ to form inputs for the trading strategy};$ 
13 for stock in stocks
14     if  $P[i]$  satisfies emotion_thresholds and  $sign(S[i])$  indicates clear direction then
15         if positive direction
16              $signal[i] \leftarrow \text{buy};$ 
17         else if negative direction
18              $signal[i] \leftarrow \text{sell};$ 
19         else
20              $signal[i] \leftarrow \text{hold};$ 
21         end if
22     end if
23 end for
24 apply trade_rules to translate signal into execution parameters and timing;
25 append all decisions to signals_logs;
26 output signals_log;

```

For the selected *report – month* relevant news articles should be collected, cleaned, and vectorised. A trained sentiment classifier would assign article-level sentiment labels, and an emotion scorer would produce intensity vectors. These outputs would be aggregated to the stock-month level and merged with month-end prices to form the decision inputs. Trading signals would then be generated by applying the calibrated emotion thresholds together with the direction of the

aggregated sentiment. All decisions would be recorded in the *signals_log* for implementation and audit.

Having established the strategy and its operational flow, the next step was to evaluate whether the generated signals were valid, reliable, and implementable under realistic market conditions. The following section outlines the approach used to test and evaluate the performance and robustness of the strategy.

7.8 Testing and evaluation

The effectiveness of the trading strategy was assessed during the testing and evaluation phase. It involved evaluating how well the trading signals resulting from the implementation of the strategy performed in a real-world scenario. Only stocks with a lag greater than zero were included in the analysis to ensure the availability of the predicted sentiment. The strategy, and consequently the output generated by the machine learning approach, was tested using the test split of the collected data.

To prevent data leakage and look-ahead bias, strict chronological partitioning was applied. Training and validation were conducted exclusively on historical data preceding the test period, and cross-validation folds were generated within this restricted window. Feature construction utilised only contemporaneously available information, ensuring that no future sentiment or price observations influenced model estimation. The final trading evaluation was conducted on a fully unseen holdout set, preserving temporal integrity.

The testing period spanned from January 2024 to June 2024, for which the emotion data was standardised to ensure consistency and comparability. Trading signals were then applied to this standardised data to enable the simulation of buy, hold, and sell decisions. The percentage of profit or loss that would have resulted from applying the strategy was calculated.

7.8.1 Approach

Using the text extracted from January 2024 to June 2024, a series of monthly trading signals was generated across seven channels (Lag 0 to Lag 6). Each signal transitioned from a default 'hold' state to either 'sell' or 'buy' in a staggered, time-shifted pattern, with each successive signal activating one period after the previous one. This staggered activation structure reflected a controlled progression of trading actions, designed to facilitate lag analysis and evaluate the timing accuracy and responsiveness of sentiment-driven signals.

To evaluate the responsiveness and timing accuracy of the sentiment-based trading signals, a lagged testing approach was applied. Lag 0 represented real-time decision-making and was

excluded from forecasting, as it relied on contemporaneous data required for testing. For Lag 1 and Lag 2, trading implied hypothetically buying the stock in February 2024 and March 2024, respectively, and holding the position through the end of the testing period due to the absence of subsequent sell signals. For Lag 3, a more dynamic strategy was required, with a buy in January 2024, a sell in February 2024, and a re-entry in April 2024 based on updated signal transitions. Similarly, Lag 4 involved a buy in January 2024, a sell in March 2024, and a repurchase in May 2024. Also, for Lag 5, the stock would have been bought in January 2024 and sold in April 2024, while Lag 6 implied a buy in January 2024 and a sell in May 2024, as dictated by the respective signal progression. This structure allowed for comparative lag analysis and performance evaluation across different signal response delays.

The testing framework operated on a signal-execution basis, whereby trading signals (buy, hold, or sell) were generated for each of the discussed stocks based on the predicted sentiment and corresponding emotional thresholds. These signals were mapped to different lag structures, depending on the lag duration that previously showed the strongest relationship with stock price movement during the model development phase.

7.8.2 Results

To assess the effectiveness of the trading strategy, the return of each selected stock was measured both individually and cumulatively over the six-month testing period. This evaluation would indicate whether the strategy correctly anticipated market direction. A signal was considered successful if it would have resulted in growth or profit. In addition, the cumulative return quantified the total gain or loss realised by executing all predicted signals, offering a comprehensive measure of the strategy's financial performance. The results are summarised in Table 7.16

Table 7.16: Performance of the dynamic trading strategy

Stock	Stock information				
	Lag (months)	Buying price (R)	Selling/holding price (R)	Change	Result
Hammerson PLC	2	272.40	288.20	5.80%	Growth
Redefine Properties	2	396.00	426.00	7.58%	Growth
FirstRand Limited	2	6516.00	8170.00	25.38%	Growth
Equites Property Fund	2	1240.00	1309.00	5.56%	Growth
Finbond Group Limited	6	30.00	60.00	100%	Growth
Growthpoint Properties	2	1094.00	1234.00	12.80%	Growth
Standard Bank Group	2	17709.00	22178.00	25.24%	Growth

For each stock, trading signals were generated based on sentiment-emotion thresholds and the most effective lag identified during the strategy development phase. The column titled 'Lag'

indicates the delay before executing the trading signal. The 'Buying price' reflects the stock price at the start of the testing period, while the 'Selling/holding price' represents the stock price at the end. Price movement over the six months is indicated in the column titled 'Change', followed by the predicted resulting outcomes in the last column. Throughout the six-month testing window, the strategy yielded favourable outcomes for every stock included in the analysis. Finbond Group Limited demonstrated the highest hypothetical return at 100%, followed by FirstRand Limited and Standard Bank Group, each exceeding a 25% growth. The average growth across all stocks was 26.07%.

The results of the study demonstrated that the developed strategy effectively carried out multiple essential tasks, including pre-processing and text cleaning. This foundational step ensured that the data used for analysis was of high quality and free from noise or irrelevant information. Following this, effective sentiment and emotion analyses were performed, which involved evaluating and interpreting the emotional tone behind a series of texts related to the general news in South Africa. Valuable insights were gained into market sentiment and public perception, both of which could significantly influence trading behaviours.

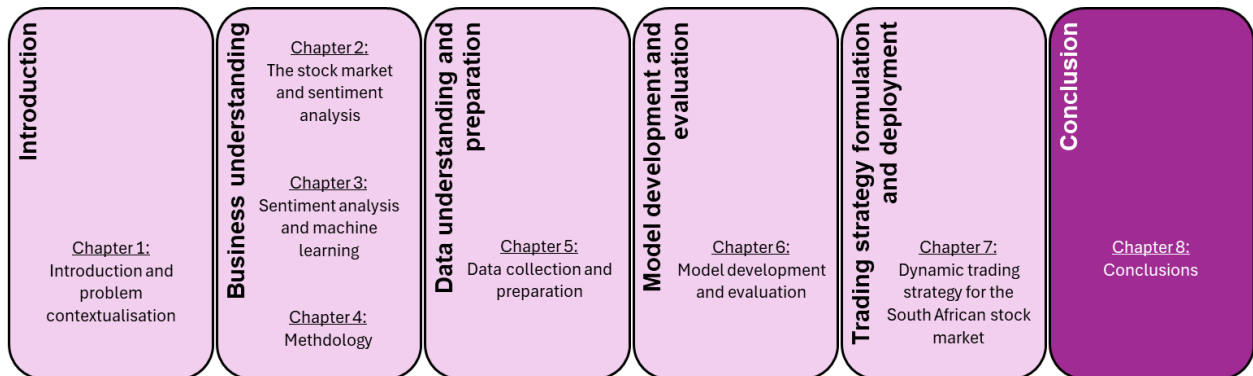
In addition, trading signals could be generated, allowing for informed decision-making regarding buy, hold, or sell actions for the selected stocks. Overall, the trading strategy demonstrated strong potential to enhance trading outcomes by using sentiment analysis and machine learning to interpret broader news signals that may influence decision-making.

7.9 Conclusion

A dynamic trading strategy based on sentiment and emotion analysis was presented in this chapter. By extracting keywords by sentiment category and mapping them to frameworks like PESTEL and Porter's Five Forces, the analysis highlighted market sentiment drivers and incorporated risk management elements. Emotion valence analysis and K-means clustering identified thresholds for emotional intensity, which helped create actionable trading signals. The strategy's effectiveness was validated through performance evaluation on a test split, demonstrating a structured approach for anticipating market movements. In the next and final chapter, a discussion on the synopsis and limitations of the study will be presented, supplemented by the contributions of the study.

CHAPTER 8

Conclusions and recommendations



8.1	Introduction	203
8.2	Synopsis of the study	203
8.3	Contributions of the study	206
8.4	Limitations and future research	208
8.5	Conclusion.....	210

8.1 Introduction

The goal of this study was to develop a machine learning approach for modelling social media sentiment to support stock market trading decisions in South Africa. In this chapter, the synopsis of the study is revisited in Section 8.2, supplemented by a systematic account of how the research objectives were met. Section 8.3 presents the contributions of this research to the relevant fields of sentiment analysis combined with machine learning in the South African stock market environment. Limitations that were identified relating to these fields are discussed in Section 8.4, together with possible directions for future research. Final remarks and conclusions are made in Section 8.5.

8.2 Synopsis of the study

This study was structured into six distinct phases, with the processes and findings documented across eight chapters. As part of the *introduction* phase, the research question was posed in the first chapter, supplemented with background information from recent literature. The *business understanding* phase comprised three chapters. A firm grasp of related work was demonstrated

through comprehensive literature reviews on the stock market, sentiment analysis, and machine learning, which guided the methodology for this research. In the *data understanding and preparation* phase, data collection, preparation, and sentiment analysis were discussed in depth. *Model development and evaluation* involved the optimisation and assessment of different machine learning models and selecting one for inclusion in the strategy and evaluation. This led to *trading strategy formulation and deployment*, where the model was integrated into a strategy for generating actionable trading signals tailored to the South African stock market.

The first chapter served to motivate and contextualise the study by highlighting the distinctiveness of the South African stock market, specifically, the JSE, introducing the research question: “*To what extent can a trading strategy be based on social media sentiment in the South African stock market?*” The following secondary objectives were formulated accordingly:

- *Conduct* a detailed literature study on the stock market and sentiment.
- *Conduct* a comprehensive literature review on sentiment analysis, data extraction from social media platforms, and machine learning in stock market trading.
- *Establish* a hybrid sentiment analysis and machine learning research methodology.
- *Select* stocks eligible for use in this study.
- *Identify* reliable social media platforms suitable for ethical data collection.
- *Collect* and *process* the data along with the selected stock price history.
- *Perform* sentiment analysis on the collected data.
- *Configure* and *evaluate* machine learning models for suitability in sentiment classification and prediction.
- *Extract* and *map* keywords to strategic frameworks.
- *Verify* the model outputs using appropriate statistical techniques.
- *Formulate, test* and *evaluate* the trading strategy for the selected stocks in the South African stock market.

To gain an integrated understanding of the research environment and intended methods to incorporate into the strategy, an in-depth literature review was *conducted*. In **Chapter 2**, the structural and functional importance of stock markets was established, with particular emphasis on the JSE. This review demonstrated that stock prices were influenced by a combination of macroeconomic, microeconomic, political, regulatory, and global factors, but were also shaped by investor perceptions and expectations. Sentiment emerged as a critical driver of supply and demand dynamics, linking market fundamentals with investor behaviour.

Building on this, the *review* of sentiment analysis and machine learning in **Chapter 3** clarified how textual data could be transformed into measurable variables through natural language processing

and pre-processing techniques. It also showed how lexicon-based, machine learning, and hybrid approaches each have distinct advantages.

Resulting from the review of related work, a hybrid methodology was *established*, which was described in detail in **Chapter 4**. This entailed discussions on data collection, pre-processing, sentiment and machine learning model selection and validation, formalisation of the trading strategy, and testing and evaluation.

From the literature, it was ascertained that stock *selection* had to be guided by liquidity, volatility, and sensitivity to sentiment. Accordingly, seven stocks from the JSE were selected as they combined sufficient historical data with the attributes most conducive to predictive modelling. The process was discussed in **Chapter 5** and supplemented by the *identification* of social media platforms suitable for the ethical collection of news headlines from South Africa. After assessing quality, availability, and ethical considerations across multiple platforms, RSS feeds and web mining of South African newspaper archives were chosen as the most viable sources of sentiment data. This dual approach ensured both breadth and temporal coverage. Detailed descriptions of the data *collection* and *pre-processing* approaches - which included preparation, feature extraction, and label assignment - were provided.

In **Chapter 6**, the impact of pre-processing techniques and sentiment distribution was discussed across different thresholding scenarios. The *performance* of the basic natural language processing sentiment analysis tool, TextBlob, and the lexicon- and rule-based tool VADER were evaluated for six shallow-learning models and two deep-learning models. Consequently, the logistic regression, random forests, and LSTM models were *configured* and *evaluated* for further investigation. Following comprehensive hyperparameter optimisation, the LSTM model achieved the highest predictive accuracy among the three models when evaluated on the unique dataset.

Chapter 7 opened with an overview of how the *extracted* keywords from the processed dataset were *mapped* to the PESTEL strategic macro-environmental forces and Porter's Five Forces framework for micro-environmental forces. This, together with a stock-specific relationship analysis, served to verify that the extracted keywords accurately reflected the underlying sentiments. To *verify* the emotional valence, emotion analysis followed, after which k-means clustering served to determine the thresholds for final trading signal generation. Accordingly, the dynamic trading strategy for selected stocks in the South African stock market was *formulated* in the form of a UML diagram, which was supplemented by an algorithm. *Testing* and *evaluation* of the strategy involved implementing the strategy on the testing split of the data. The results confirmed that the integrated data collection, pre-processing, and cleaning processes - alongside

the selected sentiment analysis and machine learning techniques - reliably generated financial growth on the test split.

With the achievement of all the research objectives, the aim of this study was fulfilled, enabling the central research question to be addressed. The findings demonstrated that social media sentiment can be effectively incorporated into a trading strategy, and, when combined with appropriate pre-processing, sentiment analysis models, and machine learning techniques, that it could yield consistently positive financial outcomes. This confirmed that sentiment-derived signals could offer meaningful predictive value within the South African stock market context.

The next section outlines the key contributions of this research, highlighting both the methodological advances and the practical implications for sentiment-driven trading, combined with machine learning, in emerging markets.

8.3 Contributions of the study

This research offered significant contributions to the field of sentiment analysis and machine learning in stock market trading, particularly within the South African context. The contributions were methodological, empirical, and practical, and together they extended the understanding of how sentiment could be harnessed within machine learning to inform trading decisions in emerging financial markets.

A dynamic trading strategy tailored to the South African stock market was developed. By translating sentiment and emotion into buy, hold, and sell signals, the research bridges the gap between theoretical modelling and actionable financial decision-making. This represents the first sentiment-driven trading strategy specifically designed and tested for the South African market.

Typically, sentiment analysis relies on predefined keywords to extract meaningful signals from the text. In contrast, this study adopted an inverse approach, correlating market movements with news headlines to identify the most relevant keywords. Although similar approaches have been explored in other research, the results of this study confirmed that this method was highly effective, providing a reliable way to link textual sentiment with market behaviour in the South African stock market. Notably, clear emotion-sentiment associations could be identified. Negative sentiment was strongly tied to anger, fear, and sadness, while positive sentiment was associated with joy and anticipation. These associations were validated through clustering methods that revealed consistent behavioural patterns, providing a more granular understanding of how emotions underpin market sentiment. The derivation of thresholds for emotional intensities using statistical boundary methods, represents another contribution, as it operationalises sentiment into quantifiable decision rules for trading.

By integrating emotion detection, cluster-based thresholding, and performance validation, a resilient methodological framework for transforming sentiment and emotion analysis into a trading strategy tailored to the South African market context was established. The deployment of this machine learning-based sentiment analysis model followed a structured approach beginning with the collection of online news headlines, followed by text pre-processing, feature extraction, and text standardisation. The processed headlines were analysed using the LSTM model for sentiment classification, while emotion analysis quantified emotional responses. These standardised emotion scores were evaluated against cluster-derived thresholds, where positive emotions (e.g., joy, trust, anticipation) corresponded to 'sell' signals and negative emotions (e.g., fear, sadness) aligned with 'buy' signals. 'Hold' signals were generated when sentiment was neutral, or when threshold conditions were not met. This systematic integration of sentiment and emotion analysis enabled practical, sentiment-informed trading decisions based on linguistic and emotional cues, demonstrating the applicability of the model of dynamic South African market conditions.

Another key contribution lay in the development of a hybrid sentiment analysis methodology that integrated lexicon-based techniques with machine learning models. By combining rule-based sentiment scoring tools such as TextBlob and VADER with supervised classifiers including logistic regression, random forests, and LSTM networks, it was demonstrated how different approaches could be strategically combined to overcome individual limitations. This hybrid approach ensured both interpretability and predictive power, establishing a replicable framework for sentiment-driven modelling.

The research further contributed by systematically addressing the challenges of data collection and pre-processing in sentiment analysis. By using Really Simple Syndication (RSS) feeds and web mining from South African newspaper platforms, access and ethical constraints were overcome, and it could be demonstrated how multiple data sources could be aligned for complementary purposes. RSS feeds provided a sufficiently large dataset for model training, while web mining ensured complete temporal coverage for correlation and lag analyses. This dual dataset strategy represented an innovative approach to solving the challenge of incomplete or inconsistent sentiment data.

A rigorous collection of pre-processing techniques was applied, tailored to noisy financial text data, which included noise removal, tokenisation, lemmatisation, and handling of emoticons and special characters. The evaluation of these techniques across models showed their impact on class distributions and overall model accuracy, which emphasised the importance of careful pre-processing in sentiment-based financial modelling.

Empirically, the study provided evidence of the predictive relationship between social media-derived sentiment and stock price movements in the South African stock market. By applying statistical validation techniques such as correlation, lag, and price position analyses, it could be determined that sentiment exerted a measurable influence on stock prices, particularly in the post-COVID-19 period when investor perceptions became more prominent in shaping market outcomes. Stock-specific lag structures were identified and provided new awareness of the way sentiment impacted returns at different intervals across companies.

The empirical findings on model performance validated the applicability of machine learning in South African financial contexts. While logistic regression and random forests offered competitive results, the LSTM model proved to be the most effective, achieving high levels of accuracy, precision, recall, and F1-scores with minimal variance during cross-validation. This exhibited the robustness and generalisability of deep learning approaches for sentiment analysis in emerging markets.

The importance of stock selection criteria was emphasised, particularly that of liquidity and volatility, when implementing sentiment-based modelling. The chosen stocks not only provided suitable data for training but also demonstrated how market attributes interacted with sentiment-driven signals to influence profitability.

By embedding emotion-specific thresholds into the trading strategy, a novel dimension to trading strategy formulation was introduced. Instead of relying solely on aggregate sentiment scores, emotional variability was integrated into decision-making, which enabled more nuanced and context-sensitive trading signals. This approach enhances responsiveness to market dynamics and provides a practical tool for investors and analysts seeking to incorporate behavioural dimensions into their strategies.

Having established the effectiveness of the strategy and highlighted its novel contributions, it is important to recognise that limitations must be acknowledged. The next section, therefore, covers these limitations and outlines avenues for future research.

8.4 Limitations and future research

While this study provided a strong foundation for sentiment-driven trading strategies in the South African stock market, several limitations must be acknowledged. Although they do not undermine the findings, they highlight areas where further research can extend and refine the work undertaken here. The following limitations must be recognised:

- The study relied on RSS feeds and web mining of South African newspaper articles, which provided valuable but incomplete coverage of possible sentiment. The RSS data contained

gaps across several months, limiting its usefulness for temporal analyses, although it remained suitable for model development. Moreover, the exclusion of social media platforms such as Twitter/X and Reddit meant that the analysis focused primarily on formal or mainstream media sentiment. While newspapers offered high-quality, structured information, they might not capture the immediacy or informality of investor sentiment expressed on real-time platforms. Future research should therefore expand to include alternative or complementary data streams to provide a fuller representation of sentiment dynamics.

- Testing was conducted on a selected set of seven stocks, chosen for their liquidity, volatility, and sensitivity to sentiment. Although this selection was deliberate and justified, it nonetheless represented a small segment of the JSE. While the objectives of this study were successfully met, further testing across additional sectors and asset classes is necessary to evaluate the broader generalisability of the results. Moreover, transaction costs, liquidity constraints in live markets, and other real-world frictions could be factored into the testing simulations.
- The trading strategy was developed at a relatively aggregated temporal level, with signals generated on a lagged monthly basis. While effective in demonstrating proof of concept, this approach limits the strategy's responsiveness to rapidly evolving market conditions.
- Although sentiment thresholds provided a useful mechanism for decision-making, the absence of a risk-adjusted framework means the model does not account for individual risk preferences or the need to mitigate losses during adverse market conditions.

After discussing the limitations of the study, attention now turns to recommendations for future research. These suggestions aim to extend the current findings, address existing constraints, and explore new opportunities for further investigation.

- Generating daily or even intraday/intermittent trading signals would allow for more frequent and granular trading decisions. The integration of macro-economic indicators alongside sentiment could further enhance predictive accuracy by accounting for broader economic forces that shape investor behaviour.
- The integration of sentiment-based trading signals with risk appetite measures would enable strategies that can adapt dynamically to investor profiles and market volatility. Similarly, portfolio-level diversification strategies, such as capital allocation across sectors or asset classes, could be embedded to enhance stability and resilience in periods of heightened uncertainty.
- Testing the portability to other exchanges - particularly in emerging economies where sentiment and behavioural influences are pronounced - could prove the generalisation of the strategy. Furthermore, advances in natural language processing, particularly transformer-based architectures and real-time streaming data analysis, offer exciting opportunities for enhancing the speed and precision of sentiment classification. By adopting these tools, future

research could refine sentiment-driven trading strategies and align them more closely with the demands of modern financial markets.

- In addition, future work will examine alternative machine learning approaches to assess their comparative performance, robustness, and suitability for sentiment-informed trading, potentially uncovering models that further improve predictive accuracy and interpretability.

In conclusion, this study demonstrated the effectiveness of integrating sentiment and emotion analysis with machine learning for trading in the South African stock market. While limitations were acknowledged, the findings provided practical insights, contributed to theory, and offered a foundation for future research in this area.

8.5 Conclusion

Chapter 8 concludes this thesis by summarising the study and demonstrating how the research objectives were met. It also highlights the study's limitations and identifies opportunities for future research.

BIBLIOGRAPHY

- Abaye, D. A., Boateng, E. Y. & Otoo, J. 2020. Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing*, 8(4): 341-357.
- Abayomi-Alli, A., Abayomi-Alli, O., Abiola, O., Misra, S. & Tale, O. A. 2023. Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and TextBlob analyser. *Journal of Electrical Systems and Information Technology*, 10(5): 1-20.
- Abbas, A., Ashraf, M., Ishaq, M. Irshad, Raza, A. A. & Sharif, W. 2022. Real-time extraction and annotation of social media contents for predicting national consumer confidence index. *Journal of Policy Research*, 8(4): 292-309.
- Abbas, F., Gattal, A. & Laouar, M. R. 2018. Automatic parameter tuning of k-means algorithm for document binarization. in: association for computing machinery (ACM). *Conference proceedings*. ICSENT 2018: Proceedings of the 7th International Conference on Software Engineering and New Technologies, Hammamet, Tunisia. 1-4.
- Abbasi, M. M. & Beltiukov, A. 2019. Summarizing emotions from text using plutchik's wheel of emotions. In: Atlantis Press. *Conference proceedings*. 7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019), Izhevsk, Russian. 291-294.
- Abd, D. H., Al-Mejibli, I. S., Alwan, J. K. & Rabash, A. J. 2018. Performance evaluation of kernels in support vector machine. In: IEEE. *Conference proceedings*. 2018 1st Annual International Conference on Information and Sciences (AiCIS), Fallujah, Iraq. 96-101.
- Abdelfattah, B. A., Darwish, S. M. & Elkaffas, S. M. 2024. Enhancing the prediction of stock market movement using neutrosophic-logic-based sentiment analysis. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(1): 116-134.
- Abdelfattah, M., Elzanfaly, D. S., Galal, D., Hassan, N. & Tallent, G. 2016. A sentiment analysis tool for determining the promotional success of fashion images on Instagram. *Sustainable Vital Technologies in Engineering & Informatics*, 11(2): 66-72.
- Abdulazeez, A. M. & Jijo, B. T. 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(1): 20-28.
- Abduljabbar, Z. H., Ali, Z. A., Almufti, S. M., Sallow, A. B. & Tahir, H. 2023. Exploring the power of extreme gradient boosting algorithm in machine learning: a Review. *Academic Journal of Nawroz University*, 12(2): 320-334.

- Abdullah, A. S., Damarta, R. & Hidayat, A. 2021. The application of k-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) Twitter account service quality. *Journal of Physics: Conference Series*, 1722(1): 1-7.
- Abdullah, R., Gan, K. H., Lee, V. L. S. & Tan, T. P. 2019. Semi-supervised learning for sentiment classification using small number of labeled data. In: *Procedia Computer Science. Conference proceedings*. The Fifth Information Systems International Conference 2019, Surabaya, Indonesia. 577-584.
- Aboutabl, A., Adel, S., Hawary, M. E. & Ouf, S. 2024. A deep learning-based LSTM for stock price prediction using Twitter sentiment analysis. (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, 15(12): 207-218.
- Abraham, S. & Kannappan, M. 2018. The role of stock exchanges in corporate governance. *International Journal of Pure and Applied Mathematics*, 119(17): 1023-1033.
- Abu-gellban, H., Nguyen, H. L., Pan, Z. & Zhang, Y. 2020. Google Trends analysis of COVID-19 pandemic. In: *IEEE. Conference proceedings*. 2020 IEEE International Conference on Big Data (Big Data), Atlanta, United States of America. 3438-3446.
- Abu-Salih, B., Alhabashneh, M., Alshamaileh, Y., Alshraideh, M., Al-Shboul, B., Awajan, A. & Zhu, D. 2023. Emotion detection of social data: APIs comparative study. *Heliyon*, 9(5): 1-14.
- Ackerson, J. M., Jim, S. & Rushit, D. 2021. Applications of recurrent neural network for biometric authentication & anomaly detection. *Information*, 12(7): 1-22.
- Adam, E., Ali, K. A. & Chabalala, Y. 2023. Exploring the effect of balanced and imbalanced multi-class distribution data and sampling techniques on fruit-tree crop classification using different machine learning classifiers. *Geomatics*, 3(1): 70-92.
- Addepalli, S., Wang Y. & Zhao, Y. 2020. Recurrent neural networks and its variants in remaining useful life prediction. *ScienceDirect*, 53(3): 137-142.
- Adekoya, A. F., Nti, I. K., Weyori, B. A. 2020. Predicting stock market price movement using sentiment analysis: Evidence from Ghana. *Applied Computer Systems*, 25(1): 33-42.
- Adelegan, O. J. 2009. The derivatives market in South Africa: Lessons for Sub-Saharan African countries. *SSRN Electronic Journal*, 2009(196): 1-18.
- Adhikary, J., Baidya, B., Ganguly, K., Mitra, M., Nayek, I. & Saha, D. 2023. Stock market prediction using sentiment analysis. *International Journal of Research Publication and Reviews*, 4(6): 2721-2729.
- Adhinata, F. D., Majid, A. & Nugraha, D. 2023. Sentiment analysis on Tiktok application reviews using Natural Language Processing approach. *Journal of Embedded System Security and Intelligent System*, 4(1): 32-38.

- Afaq, K. M., Amjad, A., Chang, H. & Khan, L. 2020. Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media. *Applied Sciences*, 12(5): 1-18.
- Aftab, F., Amphawan, A., Aslam, S., Baloch, L., Bazai, S. U., Marjan, S. & Neo, T. K. 2023. A comprehensive survey on sentiment analysis techniques. *International Journal of Technology (IJTech)*, 14(6): 1288-1298.
- Aggarwal, P., Ganaie, I. A., Kaur, R., Maqbool, J. & Mittal, A. 2023. Stock prediction by integrating sentiment scores of financial news and mlp-regressor: A machine learning approach. *Procedia Computer Science*, 218(1): 1067-1078.
- Ağrali, Ö. & Aydin, Ö. 2021. Tweet classification and sentiment analysis on metaverse related messages. *Journal of Metaverse*, 1(1): 25-30.
- Agrawal, A., Rath, S. K. & Tripathy, A. 2015. Classification of sentimental reviews using machine learning techniques. *Procedia Computer Science*, 57(6): 821-829.
- Agrawal, J. G., Chourasia, V. S. & Mittra, A. K. 2013. State-of-the-art in stock prediction techniques. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2(4): 1360-1366.
- Agudelo, C. W., Dmochowski, R. R., Everaert, K., Lazar, J. M., Monaghan, T. F., Rahman, S. N. & Wein, A. J. 2021. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina*, 57(5): 1-7.
- Aguilera, A., Barrios, S., Cardinale, Y., Dongo, I., Martinez, F. & Quintero, Y. 2020. Web scraping versus Twitter API: A comparison for a credibility analysis. In: ResearchGate. *Conference proceedings. iiWAS '20: The 22nd International Conference on Information Integration and Web-based Applications & Services*, New York, United States of America. 263-273.
- Aguilera, A., Cabeza, D., Cardinale, Y., Dongo, I., Medina, S. & Robayo, G. 2016. T-CREo: A Twitter credibility analysis framework. *IEEE Access*, 4(1): 1-20.
- Ahamed, J., Ahmad, K. & Hassan, S. U. 2022. Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, 3(1): 238-248.
- Ahmad, K., Jagatheesaperumal, S. K., Khan, J. & Sohn, K. 2025. Textual variations in social media text processing applications: challenges, solutions, and trends. *Artificial Intelligence Review*, 58(89): 1-48.
- Ahmadi, A., Asgary, A., Bragazzi, N. L., Kong, J., Mellado, B., Nia, Z. M., Ogbuokiri, B., Orbinski, J. & Wu, J. 2022. Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts. *Frontiers in Public Health*, 10(1): 1-16.

- Ahmed, M., Ali, A., Anam, S. & Naeem, A. 2023. An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*, 13(1): 911-921.
- Ahmed, M., Islam, S. M. S. & Seraj, R. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8): 1-12.
- Ahmed, W., Amin, K., Hammad, M., Pławiak, P. & Semary, N. A. 2024. Enhancing machine learning-based sentiment analysis through feature extraction techniques. *PLoS ONE*, 19(2): 1-19.
- Ahsan, A., Sufyan, M. & Tanvir, M. 2016. Investor's emotional intelligence and impact on investment decision. *International Journal of Academic Research in Economics and Management Sciences*, 5(3): 12-28.
- Ahuja, P., Ahuja, R., Chug, A., Gupta, S. & Kohli, S. 2019. The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, 152(1): 341-348.
- Aichner, T., Grünfelder, M., Jengeni, D. & Maurer, O. 2021. Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking*, 24(4): 215-222.
- Aider, F. & Palomino, M. A. 2022. Evaluating the effectiveness of text pre-processing in sentiment analysis. *Applied Sciences*, 12(17): 8765-8786.
- Äijö, J., Dimic, N., Neudl, M. & Orlov, V. 2018. Investor sentiment, soccer games and stock returns. *Research in International Business and Finance*, 43(1): 90-98.
- Ajdari, J. & Hoti, M. 2023. Sentiment analysis for TikTok review using VADER sentiment and SVM model. *SEEU Review*, 18(2): 19-33.
- Akeji, A. A. A., Ambika, M., Jabasheela, L., Kumar, S. K., Mithun, T., Sakthi, U. & Walia, R. 2022. Stock price prediction using optimal network based Twitter sentiment analysis. *Intelligent Automation & Soft Computing*, 33(2): 1217-1227.
- Akhila, A. M. S., Devi, B. S. K., Gayathri, C. & Srinivas, B. 2022. A review on sentiment analysis of Twitter data for diabetes classification and prediction. In: IEEE. *Conference proceedings. 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Solan, India. 693-698.
- Akhtar, N., Faizan, M., Mushtaq, A. & Raza, H. 2019. Scientific text sentiment analysis using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 10(12): 157-165.
- Akyüz, S. O., Atas, P. K. & Benkhaldoun, A. 2024. Predicting stock market by sentiment analysis and deep learning. *Operations Research and Decisions*, 34(2): 85-107.

- Al-Absy, M. S. M., Chalissery, N., Nishad, T. M. & Tabash, M. I. 2024. Market shocks and stock volatility: Evidence from emerging and developed markets. *International Journal of Financial Studies*, 12(1): 1-18.
- Alaghband, G. & McCarthy, S. 2023. Enhancing financial market analysis and prediction with Emotion Corpora and News Co-Occurrence Network. *Journal of Risk and Financial Management*, 16(4): 1-19.
- Al-Alawi, S. M. 2002. Development and application of back-propagation-based artificial neural network models in solving engineering. *Science and Technology*, 7(1): 55-70.
- Alam, M. M., Khan, T. A., Sadiq, R., Shahid, Z. & Su'ud, M. B. M. 2024. Sentiment analysis using support vector machine and random forest. *Journal of Informatics and Web Engineering*, 3(1): 67-75.
- Alam, S., Miah, R. & Karim, A. 2016. Analysis on factors that affect stock prices: A study on listed cement companies at Dhaka Stock Exchange. *Research Journal of Finance and Accounting*, 7(18): 2222-2847.
- Al-Amidie, M., Al-Dujaili, A., Al-Shamma, O., Alzubaidi, L., Duan, Y., Fadhel, M. A., Farhan, L., Humaidi, A. J. & Zhang, J. 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53): 1-74.
- Alamsyah, A., Az-Zahra, F., Barokah, M. R. & Trianasari, N. 2021. The correlation analysis of stock price towards public sentiment in automotive industry. In: IEEE. *Conference proceedings*. 2021 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia. 257-262.
- Albaldawi, T. H. K. & Alnuaimi, A. F. A. H. 2024. Concepts of statistical learning and classification in machine learning: An overview. *BIO Web of Conferences*, 97(12): 1-21.
- Alfeilat, H. A., Alhasanat, M. B., Eyal-Salman, H., Hassanat, A., Lasassmeh, O., Prasath, S. & Tarawneh, A. S. 2019. Effects of distance measure choice on k-nearest neighbor classifier performance: A review, *Big Data*, 7(4): 1-9.
- Alfred, R. & Chekima, K. 2016. Non-english sentiment dictionary construction. *Advanced Science Letters*, 24(2): 1416-1420.
- Al-Hameed, K. A. A. 2022. Spearman's correlation coefficient in statistical analysis. *International Journal Of Nonlinear Analysis And Applications*, 13(1): 3249-3255.
- Ali, M., Naqi, S. M. A. & Habib, M. 2023. Emotional intelligence: Understanding, assessing, and cultivating the key to personal and professional success. *Sir Syed Journal of Education & Social Research (SJESR)*, 6(2): 50-55.

- Ali, N., Hamarashid, H. K., Hassan, B., Saeed, S. & Qader, S. M. 2022. Machine learning algorithms evaluation methods by utilizing R. *UKH Journal of Science and Engineering*, 6(1): 1-11.
- AlKhatib, M. Barachi, M. E., Mathew, S. S., Nerabie, A. M. & Oroumchian, F. 2021. The impact of Arabic Part of Speech tagging on sentiment analysis: A new corpus and deep learning approach. In: Springer Link. *Conference proceedings*. The 12th International Conference on Ambient Systems, Networks and Technologies (ANT), Warsaw, Poland. 148-155.
- Almashaqbeh, M., Bakar, R. & Islam, M. A. 2021. Factors affecting share prices: A literature revisit. *AIP Conference Proceedings*, 2339(1): 1-10.
- Alnafea, M. & Chebbi, K. 2022. Does investor sentiment influence stock price crash risk? Evidence from Saudi Arabia. *Journal of Asian Finance, Economics and Business*, 9(1): 143-152.
- Aloqaily, M., Han, L., Jabber, S., Pan, Y., Rana, O. F., Yousif, M. & Zhang, D. 2020. IEEE access special section editorial: Scalable deep learning for big data. *IEEE Access*, 8(1): 216617-216622.
- Alovokpinhou, S. A. & Mbatha, V. M. 2022. The structure of the South African stock market network during COVID-19 hard lockdown. *Physica A: Statistical Mechanics and its Applications*, 590(1): 1-19.
- Aloysius, A. & Prakash, T. N. 2019. Data pre-processing in sentiment analysis using Twitter data. *International Educational Applied Research Journal (IEARJ)*, 3(7): 89-92.
- Alroobaea, R., Amin, R., Binmahfoudh, A., Hussain, M., Iqbal, A. and Iqbal, I. 2022. Sentiment analysis of consumer reviews using deep learning. *Artificial Intelligence and Sustainable Digital Transformation*, 14(17): 1-19.
- Alsahfi, T., Alshemaimri, B., Bukhari, A., Daud, A., Gardazi, N. M. & Malik, M. K. 2025. BERT applications in Natural Language Processing: a review. *Artificial Intelligence Review*, 58(166): 1-49.
- Al-Shabi, M. A. 2020. Evaluating the performance of the most important Lexicons used to sentiment analysis and opinions mining. *IJCSNS International Journal of Computer Science and Network Security*, 20(1): 51-75.
- Alslaity, A. & Orji, R. 2022. Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions. *Behaviour & Information Technology*, 43(1): 139-164.
- Amirkumar, M., Ayazbayev, D., Cankurt, S., Orynbeikova, K. & Talasbek, A. 2024. Comparative effectiveness of rule-based and machine learning methods in sentiment analysis of Kazakh language texts. *Scientific Journal of Astana IT University*, 17(1): 6-27.

- Amirthapriya, M. & Baboo, S. S. 2022. Sentiment analysis and automatic emotion detection analysis of Twitter using machine learning classifiers. *International Journal of Mechanical Engineering*, 7(2): 1161-1171.
- Anastasiu, D. C., Mohan, S., Mullanpudi, S., Sammeta, S. & Vijayvergia, P. 2019. Stock price prediction using news sentiment analysis. In: IEEE. *Conference proceedings*. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, United States of America. 205-208.
- Andleeb, R. & Hassan, A. 2023. Impact of investor sentiment on contemporaneous and future equity returns in emerging markets. *Sage Open*, 13(3): 1-11.
- Andriansyah, A. 2017. The real effects of primary and secondary equity markets on firm performance: Evidence from Indonesia. *International Journal of Managerial Finance*, 13(4): 397-418.
- Angkurawaranon, C., Angkurawaranon, S., Dawod, A. Y., Guo, Y., Phaphuangwittayakul, A. & Ying, F. 2022. An optimal deep learning framework for multi-type hemorrhagic lesions detection and quantification in head CT images for traumatic brain injury. *Applied Intelligence*, 52(1): 7320-7338.
- Anh, P. T. K., Nguyen, C. V., Phuoc, T. & Tam, P. H. 2024. Applying machine learning algorithms to predict the stock price trend in the stock market - The case of Vietnam. *Humanities and Social Sciences Communications*, 11(1): 1-19.
- Aničić, D., Petrović, V. & Simić, N. 2018. Stock and stock exchange as a part of financial institutions in developed countries. *Journal of Process Management New Technologies*, 6(1): 39-44.
- Anjani, R. M., Hananto, A. R., Nurhopipah, A., Rahayu, S. A. & Tahyudin, I. 2023. Sentiment analysis model development on e-money service complaints. *TEM Journal*, 12(4): 2050-2055.
- Ankita, A., Kumar, A., Mittal, S. & Sharma, I. 2023. Comparative analysis of shallow learning and deep learning. *Conference proceedings*. 2023 International Conference on Next Generation Electronics (NEleX), Vellore, India. 1-8.
- Anumala, M., Lakshmi, G. J., Malapolu, B. & Tholusuri, A. 2019. Sentiment analysis using LSTM. *International Journal of Engineering and Advanced Technology (IJEAT)*, 8(6S3): 1338-1340.
- Arablouei, R., Bao, Z., Culpepper, J. S., Dedeoglu, V., Huang, S. & Wang, T. 2024. Optimizing Data acquisition to enhance machine learning performance. *Proceedings of the VLDB Endowment*, 17(6): 1310-1323.
- Aradhya, M. & Kumar, S. 2020. Web scraping of LinkedIn. *International Journal of Creative Research Thoughts (IJCRT)*, 8(5): 2951-2956.

- Arief, M. & Samsudin, N. A. 2023. Hybrid approach with VADER and multinomial logistic regression for multiclass sentiment analysis in online customer reviews. *International Journal of Advanced Computer Science and Applications*, 14(12): 311-320.
- Arkhangelskaya, E. O. & Nikolenko, S. I. 2023. Deep learning for Natural Language Processing: A survey. *Journal of Mathematical Sciences*, 273(1): 533-582.
- Asongo, A. I., Barma, M. & Muazu, H. G. 2021. Machine learning techniques, methods and algorithms: Conceptual and practical insights. *International Journal of Engineering Research and Application*, 11(8): 55-64.
- Asopa, P., Goswami, A. K., Indolia, S. & Mishra, S. P. 2018. Conceptual understanding of convolutional neural network - A deep learning approach. *Procedia Computer Science*, 132(1): 679-688.
- Asudani, D. S., Nagweni, N. K. & Singh, P. 2023. Impact of word embedding models on text analytics in deep learning environment: A review. *Artificial Intelligence Review*, 56(1): 10345-10425.
- Athaiya, A., Sharma, S. & Sharma, S. 2020. Activation functions in neural networks. *International Journal of Engineering Applied Sciences and Technology*, 4(12): 310-316.
- Attílio, L. A. 2024. Transmission and impact of stock market shocks on the world economy. *Central Bank Review*, 24(1): 1-20.
- Atukalp, M. E. 2021. Determining the relationship between stock return and financial performance: An analysis on Turkish deposit banks. *Journal of Applied Statistics*, 48(13-15): 2643-2657.
- Avacharmal, R. 2024. Explainable AI: Bridging the gap between machine learning models and human understanding. *Journal of Informatics Education and Research*, 4(2): 1579-1589.
- Avetisyan, H., Bruna, O. & Holub, J. 2016. Emotion models for textual emotion classification. In: IOP Publishing. *Conference proceedings*. Journal of Physics: Conference Series, Berkeley, United States of America. 012063.
- Ayeni, J. A. 2022. Convolutional neural network (CNN): The architecture and applications. *Applied Journal of Physical Science*, 4(4): 42-50.
- Ayora, V., Horita, F. & Kamienski, C. A. 2021. Profiling online social network platforms: Twitter vs. Instagram. In: HICSS. *Conference proceedings*. Hawaii International Conference on System Sciences (HICSS 2021), Hawaii, United States of America. 2101-2110.
- Azrir, A. H. I. B. A., Haw, S. & Naveen, P. 2024. Sentiment analysis using machine learning models on Shopee reviews. *Journal of System and Management Sciences*, 14(2): 214-228.

- Baboolal, K. & Hosein, P. 2024. Bayes classification using an approximation to the joint probability distribution of the attributes. In: Springer. *Conference proceedings. Deep Learning Theory and Applications 2024*, Dijon, France. 47-61.
- Bafjaish, S. S., Hassan, N. & Ismail, M. 2020. Comparative analysis of Naive Bayesian techniques in health-related for classification task. *Journal of Soft Computing and Data Mining*, 1(2): 1-10.
- Bag, R., Bhattacharyya S. & Chakraborty, K. 2023. Sentiment analysis on labeled and unlabeled datasets using BERT architecture. *Soft Computing*, 1(1): 1-19.
- Bai, M. R., Devi, J. S. & Reddy, C. 2020. Newspaper article classification using machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(5): 872-877.
- Bakal, J. & Shah, S. S. 2022. Opinion mining for text data: An overview. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 10(6): 4301-4311.
- Balakrishnan, H. N., Kathpalia, A., Nagaraj, N. & Saha, S. 2019. ChaosNet: A chaos based artificial neural network architecture for classification. *Chaos*, 29(11): 1-18.
- Banda, K., Hall, J. H. & Pradhan, R. P. 2019. The impact of macroeconomic variables on industrial shares listed on the Johannesburg Stock Exchange. *Macroeconomics and Finance in Emerging Market Economies*, 12(3): 270-292.
- Bansal, M., Choudhary, A. & Goyal, A. 2022. A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, 3(1): 1-21.
- Barbaglia, L., Consoli, S. & Manzan, S. 2022. Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems*, 247(13): 108781-108796.
- Barnes, L., Brown, D., Heidarysafa, M., Kowsari, K., Meimandi, K. J. & Mendu, S. 2019. Text classification algorithms: A Survey. *Information*, 10(4): 150-218.
- Barz, B., Brigato, L., Denzler, J. 2022. image classification with small datasets: Overview and benchmark. *IEEE Access*, 10(1-1): 49233-49250.
- Bashar, K. 2022. A hybrid approach to explore public sentiments on COVID-19. *SN Computer Science*, 3(220): 1-19.
- Basra, J. Q. 2024. Cryptocurrency and blockchain: A comprehensive overview. *Global Media Journal*, 22(71): 1-3.
- Bassett, D. S. & Bertolero, M.A. 2020. Deep neural networks carve the brain at its joints. *Neurons and Cognition*, 1(1): 1-22.

- Basuki, A. 2023. Sentiment analysis of customers' review on delivery service provider on Twitter using Naive Bayes classification. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, 9(2): 420-428.
- Bechmann, A. & Lomborg, S. 2014. Using APIs for data collection on social media. *The Information Society*, 30(4): 256-265.
- Becker, C. 2024. Attractiveness of African stock markets for foreign investors: An analytical perspective. *Journal of Securities Operations & Custody*, 16(4): 385-395.
- Begum, M., Ferdush, J., Foysal, M. R., Islam, T., Sadik, R., Sheakh, A. & Tahosin, S. 2024. Lexicon and deep learning-based approaches in sentiment analysis on short texts. *Journal of Computer and Communications*, 12(1): 11-34.
- Belanger, A., Burden, D. & Dalfonso, P. 2022. Not all data is good data: The challenges of using machine learning with ILI. In: The American Society of Mechanical Engineers. *Conference proceedings*. Proceedings of the 2022 14th International Pipeline Conference, Calgary, Canada. 1-9.
- Belkacem, S. 2021. *Machine learning approaches to rank news feed updates on social media*. Algeria: University Of Science And Technology Houari Boumediene. (Thesis - PhD).
- Benbrahim, H. & Mercha, E. M. 2023. Machine learning and deep learning for sentiment analysis across languages: A survey. *Neurocomputing*, 531(1): 195-216.
- Benitez, J., Li, K., Liao, Q., Luo, X. R & Zhou, C. 2022. Impact of information timeliness and richness on public engagement on social media during COVID-19 pandemic: An empirical investigation based on NLP and machine learning. *Decision Support Systems*, 162(11): 1-13.
- Benjelloun, H. 2017. Evans and Archer - forty years later. *Investment management & financial innovations*, 7(1): 98-104.
- Benston, G. J. 1973. Required disclosure and the stock market: An evaluation of the Securities Exchange Act of 1934. *The American Economic Review*, 63(1): 132-155.
- Bethel, C., Cosby, A., Lebakula, V., Porter, B., Ray, T. & Stubbs-Richardson, M. 2025. A structural equation modeling approach to leveraging the power of extant sentiment analysis tools. *Journal of Computational Social Science*, 8(3): 1-21.
- Bez, J. L., Byna, S. & Hiniduma, K. 2025. Data readiness for AI: A 360-degree survey. *ACM Computing Surveys*, 57(9): 1-39.
- Bezemer, C., Li, H., Lin, D. & Rajbahadur, G. K. 2024. Keeping deep learning models in check: A history-based approach to mitigate overfitting. *IEEE Access*, 11(1): 1-14.
- Bhalla, R. & Pooja, P. 2022. A review paper on the role of sentiment analysis in quality education. *SN Computer Science*, 3(469): 1-9.

- Bhambhu, L. & Srivastava, D. 2010. Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology*, 12(1): 1-7.
- Bharathi, S. & Geetha, A. 2017. Sentiment analysis for effective stock market prediction. *International Journal of Intelligent Engineering and Systems*, 10(3): 146-154.
- Bhat, M., Nayak, S., Rao, B. A. & Reddy, N. V. S. 2022. Study of distance metrics on k-nearest neighbor algorithm for star categorization. *Journal of Physics: Conference Series*, 2161(1): 1-8.
- Bhatt, N. T. & Swarndeeep, S. J. 2020. Sentiment analysis using machine learning technique: A literature survey. *International Research Journal of Engineering and Technology (IRJET)*, 7(2): 798-802.
- Bhatti, A. A., Khokbar, M., Raza, A., Shahani, A. K. & Zehra, N. 2023. Risk assessment and decision making in stock exchange. *Journal of Social Sciences Review (JSSR)*, 3(1): 883-899.
- Bhavsar, A. K. & Gavhane, T. H. 2024. Sentiment analysis in financial markets. *International Journal of Innovative Science and Research Technology*, 9(2): 83-88.
- Diciotti, S., Espis, A. & Marzi, C. 2025. Comparative analysis of supervised and self-supervised learning with small and imbalanced medical imaging datasets. *Scientific Reports*, 15(1): 1-21.
- Biehl, M., Htun, H. H. & Petkov, N. 2023. Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, 9(26): 1-25.
- Bielikova, M., Pecar, S. & Simko, M. sentiment analysis of customer reviews: Impact of text pre-processing. In: IEEE. *Conference proceedings*. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Košice, Slovakia. 251-256.
- Biswal, B. N., Nayak, S. S. & Padhi, B. K. 2018. Machine learning for big data processing: A literature review. *International Journal of Innovative Research in Technology*, 5(7): 1-9.
- Biswas, S. K. & Bordoloi, M. 2023. Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, 56(1): 12505-12560.
- Blank, G., Dutton, W. H. & Newman, N. 2012. Social media in the changing ecology of news: The fourth and fifth estates in Britain. *International Journal of Internet Science*, 7(1): 6-22.
- Boegershausen, J., Borah, A., Datta, H. & Stephen, A. T. 2022. Fields of gold: Scraping web data for marketing insights. *American Marketing Association*, 86(5): 1-20.
- Bolek, C. & Bolek, M. 2024. Covid-19 data manipulation and reaction of stock markets. *Asia-Pacific Financial Markets*, 31(1): 137-164.
- Bonta, V., Janardhan, N. & Kumares, N. 2019. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(2): 1-6.

- Boudiaf, I. A., Frieden, I. & Scheicher, M. 2024. The market liquidity of interest rate swaps. *The Journal of Financial Market Infrastructures*, 147(1): 1-51.
- Boukes, M., van Atteveldt, W. & van der Velden, M. A. C. G. 2021. The validity of sentiment analysis: comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2): 121-140.
- Bouras, C., Pouloupoulos, V. & Tsogkas. 2010. Adaptation of RSS feeds based on the user profile and on the end device. *Journal of Network and Computer Applications*, 33(4): 410-421.
- Bourgeois, J. A., Gu, S., Hug, J. H., Patel N. P. & Wang, F. 2019. A model for basic emotions using observations of behavior in drosophila. *Frontiers in Psychology*, 10(781): 1-13.
- Boyer, J. A. C. 2009. Has the 1987 crash changed the psyche of the stock market? *Review of Accounting and Finance*, 8(2): 138-154.
- Brahmbhatt, Y., Cherry, C. & Mohamed, W. 2023. Using FinBERT as a refined approach to measuring impression management in corporate reports during a crisis. *Communicare: Journal for Communication Studies in Africa*, 42(1): 64-80.
- Branko, Š., Erik, S., Minfei, L., Yidong, G., Ze, C. & Zhi, W. 2022. Microstructure-informed deep convolutional neural network for predicting short-term creep modulus of cement paste. *Cement and Concrete Research*, 152(2): 1-17.
- Brijlal, P. & Enow, S. T. 2016. Determinants of share prices: the case of listed firms on Johannesburg Stock Exchange. *Journal of Accounting and Management*, 6(1): 85-92.
- Brito-Loeza, C., Brito-Pacheco, C. & Martin-Gonzalez, A. 2020. A regularized logistic regression based model for supervised learning. *Journal of Algorithms & Computational Technology*, 14(1): 1-9.
- Britten, J. & McKane, G. 2018. Liquidity and size effects on the Johannesburg Stock Exchange (JSE). *Investment Analysts Journal*, 47(3): 229-242.
- Brzeszczyński, J., Bwanya, P. R., Charteris, A. & Szczygielski, J. J. 2024. Google search trends and stock markets: Sentiment, attention or uncertainty? *International Review of Financial Analysis*, 91(1): 1-27.
- Budi, I., Dewi, T. B. T., Indrawan, N. A., Putra, P. K. & Santoso, A. B. 2020. Community understanding of the importance of social distancing using sentiment analysis in Twitter. In: IEEE. *Conference proceedings. 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, Yogyakarta, Indonesia. 336-341.
- Burri, V. R. & Mukku, L. 2024. Stock market prediction using machine learning techniques. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, 10(2): 757-760.

- Burzykowski, T., Geubbelmans, M., Rousseau, A. & Valkenborg, D. 2023. Validation of machine learning algorithms. *American Journal of Orthodontics and Dentofacial Orthopaedics*, 164(2): 295-297.
- Butcher, B., Cook, D., Hron, J., Weller, A. & Zilka, M. 2024. Optimising human-machine collaboration for efficient high-precision information extraction from text documents. *ACM Journal on Responsible Computing*, 1(2): 1-27.
- Cabrera, D., Cubillos, A., Cubillos, C., Mellado, R. & Urrea, E. 2018. Affective algorithm for controlling emotional fluctuation of artificial investors in stock markets. *IEEE Access*, 6(99): 7610-7624.
- Caggese, A. & Pérez-Orive, A. 2022. How stimulative are low real interest rates for intangible capital. *European Economic Review*, 142(1): 1-31.
- Cakici, N. & Zaremba, A. 2024. What drives stock returns across countries? Insights from machine learning models. *International Review of Financial Analysis*, 96(1): 1-27.
- Calvo, B., Lozano, J. A. & Sánchez-Ferrera, A. 2025. A review on self-supervised learning in time series anomaly detection: Recent advances and open challenges. *ACM Computing Surveys*, 58(5): 1-35.
- Cambria, E., Dashtipour, K., Gelbukh, A., Hamalah, A. Y. A., Hussain, A., Poria, S. & Zhou, Q. 2016. Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive Computation*, 8(4): 757-771.
- Campbell, G. & Rogers, M. 2017. Integration between the London and New York Stock Exchanges, 1825-1925. *Economic History Review*, 70(4): 1185-1218.
- Campbell, J. Y., Ramadorai, T. & Ranish, B. 2019. Do the Rich Get richer in the stock market? Evidence from India. *American Economic Review: Insights*, 1(2): 225-240.
- Camps, G., Feskens, E. J. M., Kirk, D., Kok, E., Tekinerdogan, B. & Tufano, M. 2022. Machine learning in nutrition research. *Advances in Nutrition*, 13(6): 2573- 2589.
- Cao, S., Hu, C., Xie, S. & Zhang, C. 2021. Research on the application of decision tree and random forest algorithm in the main transformer fault evaluation. *Journal of Physics: Conference Series*, 1732(1): 1-7.
- Cao, W., Dong, X., Ma, Q., Shi, Y. & Yu, Z. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2): 241-258.
- Carlson, C. R. & Cousineau, L. 2020. Are you sure you want to view this community? Exploring the ethics of reddit's quarantine practice. *Journal of Media Ethics*, 35(4): 1-12.
- Carvalho, M. & Keivandarian, N. 2023. A survey on sentiment classification methods and challenges. *The International FLAIRS Conference Proceedings*, 36(1): 1-3.,

- Casanez-Ventura, A., Castejón-Mateos, F., Cuenca-Jiménez, P. & Rodríguez-Ibáñez, M. 2023. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223(1): 1-14.
- Cebrián, E. & Domenech, J. 2023. Is Google Trends a quality data source? *Applied Economics Letters*, 30(6): 811-815.
- Cerri, R., De Carvalho, A. C. P. L. F., Horváth, T., Junior, S. B., Mantovani, R. G., Rossi, A. L. D. & Vanschoren, J. 2024. Better trees: An empirical study on hyperparameter tuning of classification decision tree induction algorithms. *Data Mining and Knowledge Discovery*, 38(3): 1364-1416.
- Çetinkaya-Rundel, M. & Dogucu, M. 2021. Web scraping in the statistics and data science curriculum: Challenges and opportunities. *Journal of Statistics and Data Science Education*, 29(1): 112-122.
- Chakrabarti, S., Chatterjee, T., Das, N. & Sadhukhan, B. 2022. Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak. *Social Network Analysis and Mining*, 12(92): 1-22.
- Chakraborty, M., Lalwani, V. & Sharma, U. 2019. Investor reaction to extreme price shocks in stock markets: A cross country examination. *IIMB Management Review*, 31(3): 258-266.
- Chalkias, I., Karapiperis, D., Tjortjis, C. & Tzafilkou, K. 2023. Learning Analytics on YouTube Educational Videos: Exploring sentiment analysis methods and topic clustering. *Electronics*, 12(18): 3949-3962.
- Chaminuka, P., Mushunje, A., Ngarava, S. & Zhou, L. 2022. Impact of the COVID-19 pandemic on the South African tobacco and alcohol industries: Experiences from British American Tobacco and Distell Group Limited. *Physics and Chemistry of the Earth*, 127(103186): 103186.
- Chandak, S., Kad, S., Mehta, G. & Sutar, R. 2023. Research paper on sarcasm detection. *International Journal of Creative Research Thoughts (IJCRT)*, 11(5): 596-602.
- Chandra, A., Dewi, L. & Meiliana. 2019. Social media web scraping using social media developers API and regex. *Procedia Computer Science*, 157(12): 444-449.
- Chandrashekar, R., Chittedi, K. R., Sakthivel, P. & Sampath, T. 2018. Macroeconomic variables and stock prices in emerging economies: A panel analysis. *Theoretical and Applied Economics*, 25(3): 91-100.
- Chang, Z. & Zhang, Z. 2023. Judging stock trends according to the sentiments of stock comments in expert forums. *Electronics*, 12(3): 1-19.

- Chaniyara, C., Jain, B., Jain, H. & Jakhotiya, A. 2022. Text pre-processing techniques in Natural Language Processing: A review. *International Research Journal of Engineering and Technology (IRJET)*, 9(2): 878-880.
- Chaudhari, P. 2020. Redundant data normalization using the novel data mining algorithms. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4): 6675-6683.
- Chaudhary, R. & Vaid, A. J. 2022. Review paper on impact of behavioral biases in financial decision-making. *World Journal of Advanced Research and Reviews*, 16(02): 989-997.
- Chellaswamy, K. P., Natchimuthu, N. & Faniband, M. 2020. Stock market sensitivity to macroeconomic factors: Evidence from China and India. *Asian Economic and Financial Review, Asian Economic and Social Society*, 10(2): 146-159.
- Chen, D., Gu, C. & Stan, R. 2021. Investor sentiment and the market reaction to macroeconomic news. *Journal of Futures Markets*, 41(1): 1412– 1426.
- Chen, G., Islam, M. & Jin, S. 2019. An overview of neural network. *American Journal of Neural Networks and Applications*, 5(1): 7-11.
- Chen, S., Dai, W., Liu, J., Wu, H., Xu, Z. & Zhang, Y. 2022. Digging into primary financial market: The issues of primary financial market issuance and investigations from the perspective of blockchain. *Frontiers in Blockchain*, 5(1): 1-17.
- Chen, Y. & Yao, J. 2021. Sentiment analysis using Part-of-Speech-based feature extraction and game-theoretic rough sets. In: IEEE. *Conference proceedings. 2021 International Conference on Data Mining Workshops (ICDMW)*, Auckland, New Zealand. 110-117.
- Chen, Z., Lien, D. & Lin, Y. 2021. Sentiment: The bridge between financial markets and macroeconomy. *Journal of Economic Behavior & Organization*, 188(5): 1177-1190.
- Cheng, J., Huang, F., Li, J. & Shi, J. 2012. Brief introduction of back propagation (bp) neural network algorithm and its improvement. *Advances in Intelligent and Soft Computing*, 169(1): 553-558.
- Cheng, K., Song, Z. & Yue, Y. 2020. Sentiment classification based on Part-of-Speech and self-attention mechanism. *IEEE Access*, 8(1): 16387-16396.
- Cherfi, A., Ferchichi, A. & Nourira, K. 2020. MC4.5 decision tree algorithm: An improved use of continuous attributes. *International Journal of Computational Intelligence Studies*, 9(1/2): 4.
- Chicco, D. & Jurman, G. 2023. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(4): 1-23.
- Chiche, A. & Yitagesu, B. 2022. Part of Speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(10): 1-25.

- Chie, D., Ho, K., Li, S. & Zerong, X. 2023. Sentiment analysis and topic modeling regarding online classes on the Reddit platform: Educators versus learners. *Applied Sciences*, 13(4): 2250-2267.
- Chifu, A. & Fournier, S. 2023. Sentiment difficulty in aspect-based sentiment analysis. *Mathematics*, 11(22): 4647-4690.
- Chitundu, J. T. 2022. A study of the factors contributing to the development of Tanzania derivatives markets. *Global Journal of Economics and Trade*, 1(1): 52-64.
- Choudhry, I. A., Iqbal, S., & Li, J., Mahmood, T. & Qureshi, A. N. 2023. Dynamic learning for imbalanced data in learning chest X-ray and CT images. *Heliyon*, 9(1): 1-20.
- Christiansen, H. & Koldertsova, A. 2009. The role of stock exchanges in corporate governance. *OECD Journal: Financial Market Trends*, 2009(1): 209-238.
- Christophides, V., Du Mouza, C., Hmedeh, Z., Scholl, M., Travers, N. & Vouzoukidou, N. 2011. Everything you would like to know about RSS feeds and you are afraid to ask. In: HAL Open Science. *Conference proceedings*. BDA'11, Base de Données Avancées, Rabat, Morocco. 1-20.
- Chua, T., Fei, H., Ji, D., Li, C., Ren, Y. & Zhang, M. 2022. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2): 1-32.
- Chue, T. K., Gul, F. A. & Mian G. M. 2019. Aggregate investor sentiment and stock return synchronicity. *Journal of Banking and Finance*, 108(1): 1-67.
- Chung, K. H. & Lee, J. L. 2016. High-frequency trading: Review of the literature and regulatory initiatives around the World. *Asia-Pacific Journal of Financial Studies*, 45(1): 7-33.
- Chung, R., Darrat, A. F. & Li, B. 2013. Seasonal anomalies: a closer look at the Johannesburg Stock Exchange. *Contemporary Management Research*, 9(2): 155-168.
- Clark, R., Evans, J., Gordon, P., Lappeman, J. & Sierra-Rubia, L. 2020. Studying social media sentiment using human validated analysis. *MethodsX*, 7(1): 1-6.
- Collada-Pérez, S., González-Cristóbal, J. C., Lana-Serrano, S. & Villena-Román, J. 2011. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In: Flairs. *Conference proceedings*. Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, Palm Beach, Florida, United States of America. 323-328.
- Cole, E., Defraeye, T., Knott, M., Kondapaneni, N., Marks, M., Perez-Cruz, F. & Pietro, P. 2025. A closer look at benchmarking self-supervised pre-training with image classification. *International Journal of Computer Vision*, 133(8): 5013-5025.

- Constantinescu, N. 2023. Research and cultural metadata exposed using APIs: Steps on the interoperability road. *Romanian Journal of Library and Information Science*, 19(1): 18-30.
- Cordes, H., Nolte, S. & Schneider, J. C. 2023. Dynamics of stock market developments, financial behavior, and emotions. *Journal of Banking & Finance*, 154(1): 106711.
- Cordova-Esparza, D., Ramirez-Pedraza, A., Romero-González, J. & Terven, J. 2025. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58(7): 1-172.
- Cordova-Esparza, D., Ramirez-Pedraza, A., Terven, J. & Urbiola, E.A.C. 2023. Loss functions and metrics in deep learning - A review. *Computer Science*, 1(1): 1-43.
- Cortis, K. & Davis, B. 2021. Over a decade of social opinion mining: a systematic review. *Artificial Intelligence Review*, 54(7): 4873-4965.
- Courdent, A. & McClelland, D. 2022. The impact of algorithmic trading on market quality: Evidence from the Johannesburg Stock Exchange. *Investment Analysts Journal*, 51(3): 157-171.
- Cristescu, M. P., Mara, D. A., Nerisanu, R. A. & Oprea, S. 2022. Using market news sentiment analysis for stock market prediction. *Mathematics*, 10(22): 1-12.
- Cui, N. 2018. Applying gradient descent in convolutional neural networks. *Journal of Physics: Conference Series*, 1004(1): 1-8.
- D'Andrea, A., Ferri, F., Grifoni, P. & Guzzo, T. 2015. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3): 26-33.
- Dahiya, N., Gupta, S. & Singh, S. 2022. A review paper on machine learning applications, advantages, and techniques. *ECS Transactions*, 107(1): 6137-6150.
- Dai, Q., Wu, H. & Zhang, C. 2016. Research of decision tree classification algorithm in data mining. *International Journal of Database Theory and Application*, 9(5): 1-8.
- Dalal, N., Kumar, J. & Sethi, M. 2024. Hyperparameters in deep learning: A comprehensive review. *International Journal of Intelligent Systems and Applications in Engineering*, 12(4): 4015-4023.
- Dara, S., Kaushik, B., Rao, C. S. & Wankhede, M. 2017. A sentiment analysis of food review using logistic regression. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2(7): 251-260.
- Dardana, J. A. & Elgeka, H. W. S. 2023. The role of emotional states on purchase decision-making among novice stock investors. *Gadjah Mada Journal of Psychology*, 9(2): 261-276.

- Das, R. K., Hasan, M. M., Hassan, M., Islam, M., Khushbu, S. A. & Razia, S. 2023. Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models. *Heliyon*, 9(9): 1-20.
- Dash, S. R., Debata, B. & Mahakud, J. 2020. Stock market liquidity: Implication of local and global investor sentiment. *Journal of Public Affairs*, 21(3): 1-19.
- De Beer, J., Keyser, N. & van der Merwe, I. 2015. The Johannesburg Stock Exchange (JSE) returns, political development and economic forces: A historical perspective. *Journal for Contemporary History*, 40(2): 1-24.
- De Bondt, W. 2020. Investor and market overreaction: A retrospective. *Review of Behavioral Finance*, 12(1): 11-20.
- De Gruijl, J. R, De Ruijt, C. & Jonker, A. 2020. Bag & Tag'em - A new dutch stemmer. In: European Language Resources Association (ELRA). *Conference proceedings*. Proceedings of the 12th International Conference on Language Resources and Evaluation, Marseille, France. 3868-3876.
- Deepa, J., Julian, A. & Umarani, V. 2021. Sentiment analysis using various machine learning and deep learning techniques. *Journal of the Nigerian Society of Physical Sciences*, 3(4): 385-394.
- Defit, S., Nurcahyo, G. W. & Zamsuri, A. 2023. classification of multiple emotions in Indonesian text using the k-nearest neighbor method. *Journal of Applied Engineering and Technological Science*, 4(2): 1012-1021.
- Defriani, M., Hermanto, T. I. & Sari, A. W. 2023. sentiment analysis of tourist reviews using k-nearest neighbors algorithm and support vector machine. *SinkrOn*, 8(3): 1366-1378.
- Dehanka, V., Rahate, S., Surjuse, V. R. & Teppalwar, T. 2022. Review sentimental analysis. *International Journal of Computer Science and Mobile Computing*, 11(3): 37-41.
- Dehmer, M. & Emmert-Streib, F. 2022. Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2): 1-24.
- DeNero, J., Guendelman, S., Prata, N. & Zepecki, A. 2020. Using application programming interfaces to access Google data for health research: Protocol for a methodological framework. *JMIR Research Protocols*, 9(7): 1-12.
- Deshpande, G., K, R., P., Khandagle, V. & Kolhe, J. 2023. Comparison of classification algorithm for crop decision based on environmental factors using machine learning. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(9s): 360-368.
- Deveci, M., Han, X., Parmar, M., Wang, L., Zhang, Y. & Zhao, X. 2024. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(99): 1-43.

- Devi, D., Kumar, S. & Krishnendhu, S. C. 2020. Review of sentiment analysis: A multilingual approach. *International Journal of Advanced Research in Computer and Communication Engineering*, 9(1): 53-58.
- Dewan, K., Ganguly, S., Garg, R., Kumar, A., Nayyar, A. & Raman, S. 2023. Explaining sentiment analysis results on social media texts through visualization. *Multimedia Tools and Applications*, 82(1): 22613–22629.
- Dhabarde, S. 2019. Approach towards model evaluation, model selection, and algorithm selection in machine learning. *Pramana Research Journal*, 9(6): 396-408.
- Dharma, A. S. & Saragih, Y. G. R. 2022. comparison of feature extraction methods on sentiment analysis in hotel reviews. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 6(4): 2349-2354.
- Dicle, M. F. & Levendis, J. 2013. The impact of technological improvements on developing financial markets: The case of the Johannesburg Stock Exchange. *Review of Development Finance*, 3(4): 204-213.
- Dishar, H. K. & Muhammed, L. A. 2023. A review of the overfitting problem in convolution neural network and remedy approaches. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, 15(2): 155-165.
- Do, R. K. G., Nishio, M., Togashi, K. & Yamashita, R. 2018. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*, 9(1): 611–629.
- Dobbin, K. K. & Simon, R. M. 2011. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, 4(31): 1-8.
- Dong, H., Liu, R. & Tham, A. W. 2024. Accuracy comparison between five machine learning algorithms for financial risk evaluation. *Journal of Risk and Financial Management*, 17(2): 1-19.
- Doty, J. R. The role of Securities and Exchange Commission in an internationalized marketplace. *Fordham Law Review*, 60(6): 77-90.
- Doulamis, A., Doulamis, N., Georgoulas, I., Kopsiaftis, G., Markoulidakis, I. & Rallis, I. 2021. Multiclass confusion matrix reduction method and its application on net promoter score classification problem. *Technologies*, 9(4): 1-22.
- Doyle, T. E., Heydarian, M. & Samavi, R. 2022. MLCM: Multi-label confusion matrix. *IEEE Access*, 10(1): 19083-19095.
- Drus, Z. & Khalid, H. 2019. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161(1): 707-714.
- Dsouza, A. P., Kanse, V. P., Kumar, K. & Vibha, V. 2024. macroeconomic factors and their impact on stock market volatility: Trends, drivers, and policy recommendations. *International Journal of Scientific Research in Engineering and Management*, 8(12): 1-5.

- Du, B., Hu, H., Luo, Y., Shen, L., Tang, A., Tao, D., Wen, Y. & Zheng, H. 2023. Learning from models beyond fine-tuning. *Journal of Late Class Files*, 14(8): 1-20.
- Du, W., He, X., Shen, M., Xv, C. & Zhang, X. 2018. Survey of convolutional neural network. *Advances in Intelligent Systems Research*, 147(1): 93-97.
- Dumas, M., Milani, F., Nolte, A. & Plotnikova, V. 2022. Designing a data mining process for the financial services domain. *Journal of Business Analytics*, 6(1): 1-27.
- Duong, H. & Nguyen-Thi, T. 2021. A review: Pre-processing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, 8(1): 1-16.
- Dutta, S. & Raj, S. 2023. A review on self-supervised learning. *International Journal of New Technology and Research (IJNTR)*, 9(1): 6-10.
- Duxbury, D., Su, C. & Wang, W. 2022. The conditional impact of investor sentiment in global stock markets: A two-channel examination. *Journal of Banking and Finance*, 138(10): 1-21.
- Easton, J. M., Fu, Q. & Nicholson, G. L. 2024. Understanding data quality in a data-driven industry context: Insights from the fundamentals. *Journal of Industrial Information Integration*, 42(1): 1-20.
- Ebersohn, J., Lekoma, M., Phiri, A., Tapa, N. & Tom, Z. 2017. The unemployment-stock market relationship in South Africa: Evidence from symmetric and asymmetric cointegration models. *Managing Global Transitions*, 15(3): 231–254.
- Edwards, C., Faculak, N., Karman, S. & Stoll, B. 2015. Social presence on LinkedIn: Perceived credibility and interpersonal attractiveness based on user profile picture. *Online Journal of Communication and Media Technologies*, 5(4): 102-115.
- Einstein, A. J., Liang, J. X., Pencina, M. & Singh, V. 2021. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports*, 11(1): 1-8.
- Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4): 169-200.
- Elbagir, S. & Yang, J. 2019. Twitter sentiment analysis based on ordinal regression. *IEEE Access*, 7(1): 1-11.
- Elgeldawi, E., Galal, A. R., Sayed, A. & Zaki, A. M. 2021. hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. *Informatics*, 8(79): 1-21.
- El-Wassal, K. A. 2013. The development of stock markets: In search of a theory. *International Journal of Economics and Financial Issues*, 3(3): 606-624.
- Elwirehardja, G. N., Isnani, M. & Pardamean, B. 2023. Sentiment analysis for TikTok Review using VADER sentiment and SVM model. *8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023)*, 227(12): 168-175.

- Enogwe, S. U., Eze, F. C., Olusola, M. O. & Oti, E. U. 2021. Comprehensive review of k-means clustering algorithms. *International Journal of Advances in Scientific Research and Engineering (IJASRE)*, 7(8): 64-68.
- Ergun, U. & Özlen, S. 2012. Macroeconomic factors and stock returns (ARDL approach). *International Journal of Academic Research in Business and Social Sciences*, 2(9): 315-343.
- Escobar, A., Moreno, J. & Múnera, S. 2013. A technical analysis indicator based on fuzzy logic. *Electronic Notes in Theoretical Computer Science*, 292(2): 27-37.
- Eswar, G. S., Rao, A. S., Sahithi, G. S., Srirupa, G. & Vanisri, L. 2024. sentiment analysis in financial market: Predicting stock markets trends based on social media sentiment. *International Journal of Research Publication and Reviews*, 5(11): 7141-7145.
- Eze, M. 2024. Digital marketing in emerging economies: A comparative study of consumer engagement strategies in Nigeria and South Africa. *International Journal of Marketing Studies*, 16(2): 65-91.
- Fahmy, M. M. 2023. Confusion matrix in three-class classification problems: A step by-step tutorial. *Journal of Engineering Research*, 7(1): 1-11.
- Fan, Y., Liu, L., Zhang, J. & Zhuang, L. 2016. Wireless channel propagation scenarios identification: A perspective of machine learning. *IEEE Access*, 4(1): 1-8.
- Fanani, M. I., Istiadi, N., Priyandoko, G. & Rofii, F. 2021. Modeling of convolutional neural networks for detection and classification of three vehicle classes. *Journal of Physics: Conference Series*, 1908(1): 1-7.
- Farid, D. M., Fatema-Tuj-Jahra, M., Hasan, M. T. & Islam, N. 2023. KNNTree: A new method to ameliorate k-nearest neighbour classification using decision tree. In: IEEE. *Conference proceedings*. 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh. 1-6.
- Fathalla, R. 2020. Emotional models: Types and applications. *International Journal of Synthetic Emotions*, 11(2): 1-28.
- Fathurrahman, H. I. K., Frisky, A. Z. K., Haq, Q. M. U., Ma'arif, A., Purwono, P. & Rahmianar, W. 2023. Understanding of convolutional neural network (CNN): A review. *International Journal of Robotics and Control Systems*, 2(4): 739-748.
- Ferreira, R. C., Goldman, A. & Shimaoka, A. M. 2023. The evolution of CRISP-DM for data science: Methods, processes and frameworks. *SBC Reviews on Computer Science*, 4(1): 1-16.
- Ferreira-Schenk, S., Matlhaku, K. & Moodley, F. 2025. The effects of investor sentiment on stock return indices under changing market conditions: Evidence from South Africa. *International Journal of Financial Studies*, 13(2): 1-24.

- Ferreira-Schenk, S., Lawrence, B. & Moodley, F. 2024. Macroeconomic determinants of the JSE size-base industries connectedness: evidence from changing market conditions. *Cogent Economics & Finance*, 12(1): 1-15.
- Feuerriegel, S., Heitzmann, S. F. & Neumann, D. 2015. Do investors read too much into news? How news sentiment causes price formation. In: IEEE. *Conference proceedings*. 2015 48th Hawaii International Conference on System Sciences (HICSS), Kauai, United States of America. 4803-4812.
- Fiesler, C., Gilbert, S., Jones, N., Proferes, N. & Zimmer, M. 2021. Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7(2): 1-14.
- Fiesler, C., Gilbert, S., Jones, N., Proferes, N. & Zimmer, M. 2024. Remember the human: A systematic review of ethical considerations in Reddit research. *Proceedings of the ACM on Human-Computer Interaction*, 8(5): 1-33.
- Firer, C. & Neu-Ner, M. A. 2015. The benefits of diversification on the JSE. *Investment Analysts Journal*, 26(46): 45-59.
- Fiske, S. T. & Schacter, D. L. 2023. Introduction. *Annual review of psychology*, 73(1): V.
- Forestier, G., Foumani, N. M., Miller, L., Salehi, M., Tan, C. W. & Webb, G. I. 2024. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9): 1-45.
- Forghani, R., Maleki, F., Muthukrishnan, N., Ovens, K. & Reinhold, C. 2020. Machine learning algorithm validation. *Neuroimaging Clinics of North America*, 30(4): 433-445.
- Fottner, A., Okhrin, Y., Pfahler, J. & Wustl, J. 2022. Reddit financial image post sentiment dataset. *Data in Brief*, 45(6): 1-9.
- Fuller, A. 2022. Predicting stock market indicators through sentiment analysis on Twitter, Iowa City: University of Iowa.
- Gagiano, M. & Marivate, V. 2023. Emotionally driven fake news in South Africa. *EPiC Series in Computing*, 93(1): 56-67.
- Gajapathy, V. & Sathyanarayana, K. 2022. Stock market prediction using sentiment analysis: A review. *Empirical Economics Letters*, 21(12): 173-188.
- Gao, Y., Lou, H., Wang, X. & Zeng, Q. 2022. Comparison of ARIMA model, DNN model and LSTM model in predicting disease burden of occupational pneumoconiosis in Tianjin, China. *BMC Public Health*, 22(2167): 1-15.
- Garg, N. & Sharma, K. 2022. Text pre-processing of multilingual for sentiment analysis based on social network data. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(1): 776-784.

- Gauthier, J., Hullé, M., Pierre, J. & Rispe, C. 2021. Critical windows: A method for detecting lagged variables in ecological time series. *Ecological Informatics*, 61(1): 101178.
- Geetha, V., Gomathy, C. K., Praneesh, S. & Yagn, S. D. V. Y. 2023. The role of natural language processing. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 7(11): 1-11.
- Geethangili, D. & Suresh, P. 2022. machine learning approach based sentiment analysis, classification: An application of Natural Language Processing. *Mathematical Statistician and Engineering Applications*, 71(4): 773-786.
- Geisler, T., Ray, H., Xie, Y. & Zhang, L. 2022. Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function. *Journal of Applied Statistics*, 49(13): 3257-3277.
- Gelderblom, O. & Jonker, J. 2004. Completing a financial revolution: The finance of the Dutch East India trade and the rise of the Amsterdam capital market, 1595–1612. *The Journal of Economic History*, 64(3): 641-672.
- Ghatora, P. S., Hosseini, S. E., Iqbal, M. J., Pervez, S. & Shaukat, N. 2024. Sentiment analysis of product reviews using machine learning and pre-trained LLM. *Big data and cognitive computing*, 8(12): 1-18.
- Ghodasara, K. 2021. Overview of decision tree pruning in machine learning. *International Research Journal of Engineering and Technology (IRJET)*, 8(8): 2073-2076.
- Ghosh, P., Gupta, S. & Jain, K. 2023. A hybrid model for sentiment analysis based on movie review datasets. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(5s): 424-431.
- Ghous, H., Hussain, I. & Zahra, T. 2021. Sentiment analysis of Twitter dataset using LLE and classification methods. *International Research Journal of Modernization in Engineering Technology and Science*, 3(1): 1151-1164.
- Giang, N. L., Mac, H., Tran, D., Tran, H. A. & Van, T. 2018. A LSTM based framework for handling multiclass imbalance in DGA botnet detection. *Neurocomputing*, 275(31): 2401-2413.
- Gide, E., Haque, I., Lu, H., Moni, M. A. & Uddin, S. 2022. Comparative performance analysis of k-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, 12(1): 1-12.
- Gilbert, E. & Hutto, C. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 216-225.

- Gilbert, E., Steyn, J. P. & Viviers, S. 2024. The low-volatility effect in African frontier equity markets. *Investment Analysts Journal*, 53(3): 189-206.
- Gilbertson, B. P. & Roux, F. J. P. 2006. The behaviour of share prices on the Johannesburg Stock Exchange. *Journal of Business Finance & Accounting*, 5(2): 223-232.
- Gill, N. S. & Singh, S. 2013. Analysis and study of k-means clustering algorithm. *International Journal of Engineering Research & Technology (IJERT)*, 2(7): 2546-2551.
- Gong, X., Li, Z., Xu, W. & Zhang, W. 2022. Uncertainty index and stock volatility prediction: evidence from international markets. *Financial Innovation*, 8(1): 1-44.
- Goodacre, R. & Xu, Y. 2018. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3): 249-262.
- Goutier, M., Kühl, N., Satzger, G. & Schemmer, M. 2022. Artificial intelligence and machine learning. *Fundamentals*, 32(1): 2235-2244.
- Griffith, J. 2020. Emotions in the stock market. *Journal of Behavioral Finance*, 21(1): 42-56.
- Gujjar, P. & Kumar, P. 2021. Sentiment analysis: Textblob for decision making. *International Journal of Scientific Research & Engineering Trends*, 7(2): 1097-1099.
- Gujrati, R. 2015. Microeconomic and macroeconomic: Issues & effects on economic growth. *International Journal of Recent Scientific Research*, 6(7): 5310-5317.
- Gulbak, E. & Kirci, P. 2020. Sentiment analysis with Instagram data. In: IEEE. *Conference proceedings*. 34th International Conference on Information Technologies (InfoTech-2020), Bulgaria, Europe. 75-80.
- Gunawan, S. K. V., Suseno, J. E. & Umargono, E. 2020. K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. In: Atlantis Press. *Conference proceedings*. Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019), Yogyakarta, Indonesia. 121-129.
- Guney, Y., Kallinterakis, V. & Komba, G. 2017. Herding in frontier markets: Evidence from African stock exchanges. *Journal of International Financial Markets, Institutions & Money*, 47(2): 152-175.
- Gupta, A. & Gupta, S. 2019. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, 161(16): 466-474.
- Gupta, A. 2024. A comprehensive study of stock market volatility: Types, determinants, and measurement methods. *International Journal For Multidisciplinary Research*, 6(6): 1-14.
- Gupta, S. K., Jhimel, A. I., Khan, M. M., Sutradhar, K. & Sutradhar, S. 2021. Stock market prediction using recurrent neural network's LSTM architecture. In: IEEE. *Conference*

- proceedings*. 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, USA. 0541-0547.
- Gupta, S., Jain, P., Mathur, K. & Mathur, P. 2023. Sentiment analysis framework of social media text by feature extraction and machine learning model. *Indian Journal of Science and Technology*, 16(29): 2233-2243.
- HaCohen-Kerner, Y., Miller, D. & Yigal, Y. 2020. The influence of pre-processing on text classification using a bag-of-words representation. *PLoS ONE*, 15(5): 1-22.
- Haddi, E., Lui, X. & Shi, Y. 2013. *Procedia Computer Science*, 17(220): 26-32.
- Hadinata, D. R., Indrayuni, E., Kristiyanti, D. A., Nurdani, A. & Putri, D. A. 2020. Comparison of Naive Bayes algorithm and support vector machine using PSO feature selection for sentiment analysis on E-wallet review. *Journal of Physics: Conference Series*, 1641(1): 1-6.
- Haenlein, M. & Kaplan, A. 2019. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 61(4): 5-14.
- Hajar, M., Hassan, A. M. & Yasmina, D. 2016. Using YouTube comments for text-based emotion recognition. *Procedia Computer Science*, 83(1): 292-299.
- Halgamuge, M. N., Pradha, S. & Vinh, N. T. Q. 2019. Effective text data pre-processing technique for sentiment analysis in social media data. *Conference proceedings*. 2019 11th International Conference on Knowledge and Systems Engineering (KSE), Da Nang, Vietnam. 1-8.
- Hameed, B. I. 2019. Blockchain and cryptocurrencies technology: A survey. *International Journal on Informatics Visualization*, 3(4): 355-360.
- Haneczok, J. & Piskorski, J. 2020. Shallow and deep learning for event relatedness classification. *Information Processing and Management*, 57(6): 1-25.
- Hanif, W. & Mehmood, Y. 2014. Impact of bullish and bearish market on investor sentiment. *International Journal of Innovation and Applied Studies*, 9(1): 142-151.
- Hansen, K. B. & Borch, C. 2022. Alternative data and sentiment analysis: prospecting non-standard data in machine learning-driven finance. *Big Data & Society*, 9(1): 1-14.
- Hansun, S. & Young, J. C. 2021. Predicting LQ45 financial sector indices using RNN-LSTM. *Journal of Big Data*, 8(104): 1-13.
- Hargreaves, C. A. & Ridhwan, K. M. 2021. Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore. *International Journal of Information Management Data Insights*, 1(2): 100021.
- Harnish, K. N., Monika, P., Kulkarni, C., Shruthi, S. & Vani, V. 2022. Machine learning approaches for sentiment analysis: A survey. *International Journal of Health Sciences*, 6(S4): 1286-1300.

- Hasib, K. M., Ranjan, R., Shamrat, J. M., Siddique, A. H. & Yadav, A. 2022. Performance evaluation among ID3, C4.5, and CART decision tree algorithm. In: Springer. *Conference proceedings*. Pervasive Computing and Social Networking, Singapore, Asia. 127-142.
- Hassenstein, M. J. & Vanella, P. 2023. Data quality - concepts and problems. *Encyclopedia*, 2(1): 498-510.
- Hattingh, M., Kekere, T. & Marivate, V. 2023. Exploring COVID-19 public perceptions in South Africa through sentiment analysis and topic modelling of Twitter posts. *The African Journal of Information and Communication*, 2023(31): 1-27.
- He, M. & Qu, H. 2022. Predicting volatility based on interval regression models. *Journal of Risk and Financial Management*, 15(12): 1-21.
- Heertje, A. 2004. Schumpeter and methodological individualism. *Journal of Evolutionary*, 14(2): 153-156.
- Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V. & Pérez-Meana, H. 2018. A web scraping methodology for bypassing Twitter API restrictions. *ArXiv*, 1(1): 1-7.
- Hidayati, N. N., Raharjo, A. B. & Rochimah, S. 2022. Software traceability in agile development using topic modeling. *International Journal on Advanced Science, Engineering and Information Technology*, 12(4): 1410-1420.
- Hidayatullah, A. F. & Ma'arif, M. R. 2017. Pre-processing tasks in Indonesian Twitter messages. In: IOP Publishing. *Conference proceedings*. International Conference on Computing and Applied Informatics 2016, Medan, Indonesia. 1-6.
- Hlomani, H., Joseph, S. R., Kaniwa, F., Letsholo, K. & Sedimo, K. 2016. Natural Language Processing: A review. *International Journal of Research in Engineering and Applied Sciences*, 6(3): 207-212.
- Ho, W. K. O. & Kee, T. 2025. Optimizing machine learning models for urban sciences: A comparative analysis of hyperparameter tuning methods. *Urban Science*, 9(9): 1-24.
- Hofmeyr, D. 2023. Incremental estimation of low-density separating hyperplanes for clustering large data sets. *Pattern Recognition*, 139(2): 1-21.
- Holden, M., Sarpong, P. K. & Sibanda, M. 2016. Investigating chaos on the Johannesburg Stock Exchange. *Journal of Economics and Behavioral Studies*, 8(5): 56-67.
- Hoos, H. H. & Van Engelen, J. E. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(1): 373-440.
- Hopman, C. 2007. Do supply and demand drive stock prices? *Quantitative Finance*, 7(1): 37-53.

- Hora, A. & Job, R. 2024. Availability and Usage of Platform-Specific APIs: A First Empirical Study. In: IEEE. *Conference proceedings*. 2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR), Lisbon, Portugal. 27-31.
- Hoste, V. & Jacobs, G. 2021. Fine-grained implicit sentiment in financial new: Uncovering hidden bulls and bears. *Electronics*, 20(10): 1-33.
- Hsing, Y. 2011. The stock market and macroeconomic variables in a BRICS country and policy implications. *International Journal of Economics and Financial Issues*, 1(1): 12-18.
- Humpherys, S. L. & Lazrig, I. 2022. Using machine learning sentiment analysis to evaluate learning impact. *Information Systems Education Journal (ISEDJ)*, 20(1): 13-21.
- Hussain, F. Z., Khalid, S., Mukhtar, A. & Nazir, Y. 2022. The impact of firm-specific factors on the stock prices: New insight from emerging market of Asia. *Journal of Accounting and Finance in Emerging Economies*, 8 (1), 101-112.
- Hussain, N. Y. 2024. Deep learning architectures enabling sophisticated feature extraction and representation for complex data analysis. *International Journal of Innovative Science and Research Technology*, 9(10): 2290-2300.
- Huynh, T. T., Nguyen, H. T. & Phu, D. T. 2024. Enhancing fire detection performance based on fine-tuned YOLOv10. *Computers, Materials and Continua*, 81(2): 2281-2298.
- Iglesias, C. A. & Moreno, A. 2019. Sentiment analysis for social media. *Applied Sciences*, 9(23): 1-4.
- Ilesanmi, K. D. & Olaniran, S. O. 2021. Socioeconomic impact of the COVID-19 pandemic in South Africa: Counting the cost. *Economics of Development*, 20(2): 41-50.
- Imtiaz, N., Jamali, M., Maree, M., Memon, A. G., Soomro, A. H. & Soomro, S. 2024. RSS feeds filtering from multiple sources using automated techniques of Natural Language Processing. *The Asian Bulletin of Big Data Management*, 4(1): 1-14.
- Indah, K. A. T., Indrayana, I. N. E., Prihatini, P. M., Sudiarta, I. K. G. & Sukerti, G. N. A. 2021. Feature extraction performance on classified methods for text sentiment analysis. In: iCAST. *Conference proceedings*. Proceedings of the 4th International Conference on Applied Science and Technology on Engineering Science, Kalimantan Timur, Indonesia. 1235-1243.
- Ingersoll, G. M., Lee, K. L. & Peng, C. J. 2002. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1): 3-14.
- Irving, J. 2005. Regional integration of stock exchanges in Eastern and Southern Africa: Progress and prospects. *IMF Working Papers*, 5(122): 1-41.
- Ismail, G., Swart, L., Taliep, N. & Van Niekerk, A. 2022. The converging influence of social, economic and psychological factors on public responsiveness to the COVID-19 pandemic in South Africa. *BMC Public Health*, 22(1): 1-11.

- Jadhav, S., Kate, R., Masarkar, S., Masram, S., Takale, D. G., Patil, A. & Patil, V. 2024. Sentiment analysis through the application of machine learning algorithms. *Journal of Computer Networks and Virtualization*, 2(1): 9-14.
- Jafarov, E. 2022. Technologies of data processing and cleaning, noise identification and removal at time series. *PAHTEI-Proceedings of Azerbaijan High Technical Educational Institutions*, 23(12): 73-82.
- Jaffry, S. W., Malik, M. K. & Nasar, Z. 2021. Named entity recognition and relation extraction: State of the art. *ACM Computing Surveys*, 54(1): 1-34.
- Jain, E. 2014. Impact of law of demand & supply on stock market: A study of most active BSE indices with the help of RSI. *Journal of Economics and Finance*, 5(6): 6-15.
- Jaiswal, U. C. & Singh, N. 2023. Sentiment analysis using machine learning: A comparative study. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 12(1): 1-15.
- Jamjuntr, P., Kaewyong, P. & Suanpang, P. 2021. Sentiment analysis with TextBlob package implications for tourism. *Journal of Management Information and Decision Sciences*, 24(6): 1-9.
- Jayarathna, S. & Sazzed, S. 2021. SSentiaA: A self-supervised sentiment analyzer for classification from unlabeled data. *Machine Learning with Applications*, 4(2): 1-14.
- Jebaseeli, A. N. & Kasthuri, S. 2020. An efficient decision tree algorithm for analyzing the Twitter sentiment analysis. *Journal of Critical Reviews*, 7(4): 1010-1018.
- Jha, R., Keswani, S. & Puri, V. 2024. Relationship among macroeconomic factors and stock prices: cointegration approach from the Indian stock market. *Cogent Economics & Finance*, 12(1): 1-26.
- Jhunjunwala, Y. 2024. The impact of interest rate changes on financial markets: A mathematical study. *International Journal of Scientific Research in Science and Technology*, 11(5): 157-169.
- Ji, S., Park, J. & Yi, D. 2020. Analysis of recurrent neural network and predictions. *Symmetry*, 12(4): 1-15.
- Jiang, L., Koumarelas, I. & Naumann, F. 2020. Data preparation for duplicate detection. *Journal of Data and Information Quality*, 12(3): 1-24.
- Jianqiang, Z. & Xiaolin, G. 2017. Comparison research on text pre-processing methods on Twitter sentiment analysis. *IEEE Access*, 5(1): 2870-2879.
- Jie, H. J. & Wanda, P. 2020. RunPool: A dynamic pooling layer for convolution neural network. *International Journal of Computational Intelligence Systems*, 13(1): 66-76.

- Jin, Z., Lui, Y. Yang, Y. 2020. Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, 32(1): 9713-9729.
- John, A. & Latha, T. 2023. Stock market prediction based on deep hybrid RNN model and sentiment analysis. *Journal for Control, Measurement, Electronics, Computing and Communications*, 64(4): 981-995.
- Johnson, L., Krotov, V., & Silva, L. 2020. Tutorial: Legality and ethics of web scraping. *Communications of the Association for Information Systems*, 47(1): 555-581.
- Johnson, M. A. & Kelly, A. 2021. Investigating the statistical assumptions of Naïve Bayes classifiers. In: IEEE. *Conference proceedings. 2021 55th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, United States of America. 1-6.
- Jun, Z. 2021. The development and application of support vector machine. *Journal of Physics: Conference Series*, 1748(1): 1-6.
- Kabir, M. M., Othman, Z. A. & Yaakub, M. R. 2024. Enhanced lexicon based hybrid method for slang and punctuation scoring for aspect based sentiment analysis. In: IEEE. *Conference proceedings. 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, Dhaka, Pakistan. 1418-1422.
- Kadir, R. A., Suhaidi, M. & Tiun, S. 2021. A review of feature extraction methods on machine learning. *Journal of Information System and Technology Management*, 6 (22): 51-59.
- Kaji, N., Kitsuregawa, M., Ren, Y. & Yoshinaga, N. 2014. Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. *IEICE Transactions on Information and Systems*, 97(4): 790-797.
- Kamalanabhan, T. J. & Navaneethan, C. G. 2020. Emotional intelligence vs decision making; are they complementary? If yes, how are they related? *European Journal of Education Studies*, 7(12): 392-405.
- Kandiri, J., Ondara, B. & Waithaka, S. 2022. Machine learning techniques, features, datasets, and algorithm performance parameters for sentiment analysis: A systematic review. *Open Journal for Information Technology*, 5(1): 1-16.
- Kang, W., Ratti, R. A. & Vespignani, J. 2020. Global commodity prices and global stock market volatility shocks: Effects across countries. *Journal of Asian Economics*, 71(1): 1-18.
- Kannan, S. & Ramachandran, G. 2021. Artificial intelligence and deep learning applications: A review. *Journal of Artificial Intelligence, Machine Learning and Neural Network*, 1(02): 10-13.
- Kanozia, R., Kaur, G., Kaur, H. & Singh, S. 2023. Empirical analysis of supervised and unsupervised machine learning algorithms with aspect-based sentiment analysis. *Applied Computer Systems*, 28(1): 125-136.

- Kapočiūtė-Dzikienė, J., Krilavičius, T. & Krupavičius, A. 2013. A comparison of approaches for sentiment classification on lithuanian internet comments. In: Association for Computational Linguistics. *Conference proceedings*. Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria. 2-11.
- Karanikola, A., Kotsiantis, S. & Liapis, C. M. 2023. Investigating deep stock market forecasting with sentiment analysis. *Entropy*, 25(2): 1-30.
- Katole, M. R. S. 2024. LinkedIn data analysis and predications of recruitment ratio using Power BI. *International Journal for Research in Applied Science and Engineering Technology*, 12(10):614-619.
- Kavya, P., Marimuthu, M., Nivetheni, G. G., Saagarika, S., Subavarsshini, R. T. & Velvadivu, P. 2021. Stock market analysis using MAPREDUCE and PYSPARK. *International Research Journal of Modernization in Engineering Technology and Science*, 3(7): 1785-1795.
- Keerthi, D., Laxmi, K. B., Rakshitha, C. H. & Yamini, B. 2020. Twitter sentiment analysis using VADER on Python. *Journal of emerging technologies and innovative research*, 7(5): 1025–1031.
- Keller, D. B. & Schultz, J. 2014. Word formation is aware of morpheme family size. *PLoS ONE*, 9(4): 1-6.
- Khairani, M., Siregar, A. S. R. & Siregar, Y. S. 2023. implementation of the data mining cart algorithm in the characteristic pattern of new student admissions. *Journal Of Computer Networks Architecture and High Performance Computing*, 5(1): 263-275.
- Khairnar, J. & Kinikar, M. 2013. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, 6(3): 1-6.
- Khan, Y. 2019. *A longitudinal sentiment analysis of the #FeesMustFall campaign on Twitter*. Durban: Durban University of Technology. (Thesis - PhD).
- Khard, V. & Sonawane, S. S. 2016. Sentiment analysis of Twitter data: A survey of techniques. *International Journal of Computer Applications*, 139(11): 5-15.
- Khotimah, B. K., Rohman, E. M., Satoto, B. D. & Syakur, M. A. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, 336(1): 1-7.
- Kiss, A. & Nemes, L. 2021. Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, 5(3): 375-394.
- Komalalakshmi, J. 2023. Design and development of social media sentiment analysis on customer reviews of Amazon products for business intelligence in Python. *International Journal of Data Analytics Research and Development (IJDARD)*, 1(1): 9-23.

- Kotecha, K., Mehta, P. & Pandya, S. 2021. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science*, 7(20): 1-21.
- Kotzé, E. & Senekal, B. 2018. Employing sentiment analysis for gauging perceptions of minorities in multicultural societies: An analysis of Twitter feeds on the Afrikaner community of Orania in South Africa. *The Journal for Transdisciplinary Research in Southern Africa*, 14(1): 1-11.
- Kulkarni, C., Rao, A. C. S. & Wankhade, M. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(55): 5731-5780.
- Kumalo, N. L. 2023. The effects of liquidity and solvency on South African Banks' performance. Pretoria: University Of South Africa. (Thesis - MsC).
- Kumar, V. & Nidugala, G. K. 2020. Impact of primary market on total factor productivity: A cross-country analysis. *Cogent Economics & Finance*, 8(1): 1-24.
- Kumar, Y. V., Sravanthi, Y., Srihith, I. V. D. & Srinivas, T. A. S. 2024. Stock duel: Python's play in comparative market analysis. *Journal of Advancement in Parallel Computing*, 7(1): 1-4.
- Kumari, K. & Yadav, S. 2018. Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, 4(1): 33-36.
- Lai, C., Lin, Y. & Pai, P. 2022. Using deep learning techniques in forecasting stock markets by hybrid data with multilingual sentiment analysis. *Electronics*, 11(21): 1-19.
- Lakra, N. 2024. The influence of macroeconomic factors on mergers and acquisitions and stock market performance. *International Journal For Multidisciplinary Research*, 6(4): 1-9.
- Lakshmipriya, R., Sahithya, A., Shakeel, S. & Shashank, K. 2023. Reviews analysis using Gaussian Naïve Bayes in machine learning. *International Research Journal of Engineering and Technology (IRJET)*, 10(1): 1004-1008.
- Laxkar, P. & Tomar, A. 2022. Differences of Tanh, sigmoid and ReLu activation function in neural network. *International Journal of Scientific Progress and Research (IJSPR)*, 80(196): 18-21.
- Lebepe, B. A. 2020. Market reaction to rights offer: A case of Johannesburg Stock Exchange (Doctoral dissertation, University of the Witwatersrand).
- Lee, H. C. 2024. The relationship between investors sentiment and stock liquidity. *The Korean Data Analysis Society*, 26(5): 1455-1467.
- Li, A. 2023. Reddit sentiment analysis for natural language processing. *Applied and Computational Engineering*, 5(1): 579-582.
- Li, C., Lin, H., Liu, X., Wan, S., Xu, Y. & Xu, Y. 2023. Forward looking statement, investor sentiment and stock liquidity. *Heliyon*, 9(4): 1-16.

- Li, H., Mao, H. & Wang, J. 2021. Part-of-Speech Tagging with rule-based data preprocessing and transformer. *Electronics*, 11(1): 56-70.
- Li, M. & Zhang Y. 2023. Integrating social media data and historical stock prices for predictive analysis: A reinforcement learning approach. *International Journal of Advanced Computer Science and Applications*, 14(12): 26-45.
- Liao, Q. & Poggio, T. 2018. Theory I: Deep networks and the curse of dimensionality. *Bulletin of the Polish Academy of Sciences, Technical Sciences*, 66(6): 761-773.
- Liao, Q., Mhaskar, H. & Poggio, T. 2017. When and why are deep networks better than shallow ones? In: AAAI Press. *Conference proceedings*. Proceedings of the AAAI Conference on Artificial Intelligence, California, United States of America.
- Liao, X. 2024. Sentiment analysis based on machine learning models. *Applied and Computational Engineering*, 54(1): 129-134.
- Lim, S. H. 2021. Understanding recurrent neural networks using nonequilibrium response theory. *Journal of Machine Learning Research*, 22(1): 1-48.
- Liu, F. & Yan, T. 2022. COVID-19 sentiment analysis using college subreddit data. *PLoS One*, 17(11): 1-18.
- Liu, J., Wu, K. & Zhou, M. 2023. News tone, investor sentiment, and liquidity premium. *International Review of Economics & Finance*, 84(1): 167-181.
- Liu, Q., Mao, Y. & Zhang, Y. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Computer and information sciences*, 36(4): 1-16.
- Liu, Y., Weng, T., Wu, S. & Zou, Z. 2022. S_I_LSTM: Stock price prediction based on multiple data sources and sentiment analysis. *Connection Science*, 34(1): 44-62.
- Livisha, K., Poonguzhali, R., Vinothini, S. & Waldiya, V. 2018. Sentiment analysis on LinkedIn comments. *International Journal of Engineering Research & Technology (IJERT)*, 6(7): 1-3.
- López-Cabarcos, M. Á., López-Pérez, M. L., Pérez-Pico, A. M. & Vázquez-Rodríguez, P. 2020. Investor sentiment in the theoretical field of behavioural finance. *Economic Research-Ekonomska Istraživanja*, 33(1): 2101-2119.
- Lordkipanidze, R. 2023. New understanding of law of supply and demand and perfect economic competition for effective economic growth and political systems. *Theoretical Economics Letters*, 13(5): 152-1158.
- Lu, W. & Sun, X. 2022. Implicit n-grams induced by recurrence. In: Association for Computational Linguistics. *Conference proceedings*. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States of America. 1-11.

- Lukasiewicz, M. 2017. From diamonds to gold: The making of the Johannesburg Stock Exchange, 1880-1890. *Journal of Southern African Studies*, 43(4): 1-18.
- Lutsenko, R. R. 2024. The Application Programming Interface as a method for monitoring social network data for research in behavioral economics. *Business Inform*, 8(559): 133-141.
- Mabokela, K.R. & Schlippe, T. 2022. AI for social good: Sentiment analysis to detect social challenges in South Africa. *Artificial Intelligence Research Communications in Computer and Information Science*, 1734(1): 309-322.
- Macaulay, A. & Song, W. 2023. News media, inflation, and sentiment. *AEA Papers and Proceedings*, 113(1): 172-176.
- Maghanga, M. & Quisenberry, W. 2015. Uganda Securities Exchange as a medium of trade: To what extent have the private and public sectors embraced it? *Journal of Behavioural Economics, Finance, Entrepreneurship, Accounting and Transport*, 3(3): 96-109.
- Magwedere, M. R. & Marozva, G. 2017. Macroeconomic variables, leverage, stock returns and stock return volatility. *Macroeconomics and Monetary Economics*, 13(4): 264-288.
- Mahesh, B. 2018. Machine learning algorithms - A review. *International Journal of Science and Research (IJSR)*, 9(1): 381-386.
- Mamani-Coaquira, Y. & Villanueva, E. 2024. A review on text sentiment analysis with machine learning and deep learning techniques. *IEEE Access*, 12(1): 193115-193130.
- Mamtha, D. & Srinivasan, S. 2016. Stock market volatility - Conceptual perspective through literature survey. *Mediterranean Journal of Social Sciences*, 7(1): 208-212.
- Maqsood, H., Maqsood, M., Mehmood, I., Moon, J., Rho, S. & Yasmin, S. 2022. Analyzing the stock exchange markets of EU Nations: A case study of Brexit social media sentiment. *Systems*, 10(2): 24.
- Maram, B., Monica, N., Nikhila, P., Tejasri, T. & Rao, S. M. 2020. Positivity calculation using Vader sentiment analyser. *International Journal of Academic Engineering Research (IJAER)*, 4(3): 13-17.
- Marini, R., Pugliese, R. & Regondi, S. 2021. Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Science and Management*, 4(1): 19-29.
- Marivate, V. & Terblanche, M. 2021. Towards financial sentiment analysis in a South African landscape. *ArXivLabs*, 2106.10004(1): 185-202.
- Masethe, H. D., Masethe, M. A., Ojo, S. O. & Owolawi, P. A. 2023. Sentiment analysis of GEMS application process for Sesotho sa Leboa language using NLP. In: Social Science Research Network. *Conference proceedings*. Proceedings of International Conference on Information systems and Emerging Technologies, Windhoek, Namibia. 1-13.

- Masoud, N. M. 2013. The impact of stock market performance upon economic growth. *International Journal of Economics and Financial Issues*, 3(4): 788-798.
- Mathew, A. P. & Mathew, S. 2023. Review on techniques in sentiment analysis. *International Journal of Novel Research and Development*, 8(1): 105-114.
- Matta, P., Pant, B., Sharma, D., Sharma, N. and Sharma, S. 2020. Web scraping: Applications and scraping tools. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(5): 8202-8206.
- Mazur, M., Dang, M. and Vega, M. 2021. COVID-19 and the March 2020 stock market crash. Evidence from S&P1500. *Finance Research Letters*, 38(101690): 101690.
- McIlwain, D. 2007. Pleasure in mind: Silvan Tomkins and affect in aesthetics, personality theory and culture. *Theory & Psychology*, 17(4): 499-504.
- Meghana, P. S. & Sekhar, C. 2020. A study on backpropagation in artificial neural networks. *Asia-Pacific Journal of Neural Networks and Its Applications*, 4(1): 21-28.
- Meher, S. K., Mishra, B. S. P., Pattanaik, A. & Sagnika, S. 2020. A review on multi-lingual sentiment analysis by machine learning methods. *Journal of Engineering Science and Technology Review*, 13(2): 154-166.
- Mehta, S., Singh, T. & Varsha, M. S. 2011. Macroeconomic factors and stock returns: Evidence from Taiwan. *Journal of Economics and International Finance*, 2(4): 217-227.
- Meldrum, A. & Sokolinskiy, O. 2025. The relationship between market depth and liquidity fragility in the treasury market. *Finance and Economics Discussion Series*, 1(2025-014): 1-42.
- Mena-Chalco, J. P. & Padilha, A. 2024. Navigating technical, legal, and ethical hurdles to scraping LinkedIn data for academic research. *Liinc em Revista*, 20(1): 1-25.
- Mendoza-Urdiales, R. A., Núñez-Mora, J. A., Santillán-Salgado, R. J. & Valencia-Herrera, H. 2022. Twitter sentiment analysis and influence on stock performance using transfer entropy and EGARCH methods. *Entropy*, 24(7): 1-17.
- Merino, A. & Sebastian, A. 2019. Is the Alternative Exchange achieving its objectives? A capital structure perspective. *Journal of Economic and Financial Sciences*, 12(1): 1-23.
- Mhlongo, N. Z., Oyeniyi, L. D. & Ugochukwu, C. E. 2024. Analyzing the impact of algorithmic trading on stock market behavior: A comprehensive review. *World Journal of Advanced Engineering Technology and Sciences*, 11(2): 437-453.
- Michie, R. C. 1986. The London and New York Stock Exchanges, 1850-1914. *The Journal of Economic History*, 46(1): 171-187.
- Min, G., Muriira, M. L. & Zhao, Z. 2018. Exploiting linear support vector machine for correlation-based high dimensional data classification in wireless sensor networks. *Sensors*, 18(9): 1-23.

- Mladenova, T. 2021. A feature-weighted rule for the k-nearest neighbor. In: IEEE. *Conference proceedings*. 021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey. 493-497.
- Mofokeng, T. & Moodley, F. 2025. The effects of changing macroeconomic conditions on the Commercial Bank returns in South Africa. *International Journal of Economics and Financial Issues*, 15(2): 76-90.
- Mohamad, I. B. & Usman, D. 2013. Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6(17): 3299-3303.
- Moors, A. & Scherer, K. R. 2019. The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology*, 70(1): 719-745.
- Mostofi, H. & Shafiezzad, O. 2024. Sentiment analysis of Berlin tourists' food quality perception through artificial intelligence. *Tourism and Hospitality*, 5(4): 1396-1417.
- Mpini, S. & Tchereni, B. H. M. 2020. Monetary policy shock and stock market volatility in emerging markets. *Risk Governance & Control: Financial Markets & Institutions*, 10(3): 50-61.
- Mpofu, R. T. 2011. An analysis of the South African FTSE/JSE all-share index and the macroeconomic variables during the period 2002 to 2010. *Corporate Ownership and Control*, 8(2): 322-333.
- Muguto, H. T., Muzindutsi, P. & Rupande, L. 2019. Investor sentiment and foreign financial flows: Evidence from South Africa. *Proceedings of Rijeka School of Economics*, 37(2): 473-498.
- Muller, C. & Ward, M. 2013. Style-based effects on the Johannesburg Stock Exchange: A graphical time-series approach. *Investment Analysts Journal*, 77(1): 1-16.
- Munk, M., Munková, D. & Vozár, M. 2017. Data pre-processing evaluation for text mining: Transaction/sequence model. *Procedia Computer Science*, 18(4): 1198-1207.
- Musial, K. & Kazienko, P. 2013. Social networks on the internet. *World Wide Web*, 16(1): 31-72.
- Musial, K. & Naseem, U. 2019. DICE: Deep intelligent contextual embedding for Twitter sentiment analysis. In: IEEE. *Conference proceedings*. 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia. 953-958.
- Nainggolan, R., Perangin-angin, R., Simarmata, E. & Tarigan, F. A. 2019. Improved the performance of the k-means cluster using the sum of squared error (SSE) optimized by using the elbow method. *Journal of Physics: Conference Series*, 1361(1): 1-7.
- Nandwani, P. & Verma, R. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(81): 1-19.

- Narayan, R. & Samanta, P. 2023. A machine learning approach for sentiment analysis using social media posts. *International Journal of Information Technology and Computer Science (IJITCS)*, 16(5): 23-35.
- Nath, S. 2024. Lecture 12: Neural Networks Training Algorithm. January-April. https://www.cse.iitb.ac.in/~swaprava/courses/cs217/scribes/CS217_2024_lec12.pdf Date of access: 27 October 2025.
- Nawaz, S. 2023. A comparative study of machine learning algorithms for sentiment analysis. *International Journal of Research Publication and Reviews*, 4(8): 1999-2003.
- Nazir, Z., Park, J. & Zarymkanov, T. 2022. A machine learning model selection considering tradeoffs between accuracy and interpretability. *Advances in Science Technology and Engineering Systems Journal*, 7(4): 72-78.
- Ngene, C. C., Obinnaya, O. & Ubah, V. 2021. Machine learning types and techniques. *IOSR Journal of Engineering (IOSRJEN)*, 11(12): 56-58.
- Nguyen, H. L. & Vo, D. H. 2024. Market risk spillover and the asymmetric effects of macroeconomic fundamentals on market risk across Vietnamese sectors. *Financial Innovation*, 10(95): 1-29.
- Nguyen, T. H., Shirai, K. & Velcin, J. 2015. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24): 9603-9611.
- Ngwenya, N. & Simatele, M. D. 2020. Unbundling of the green bond market in the economic hubs of Africa: Case study of Kenya, Nigeria and South Africa. *Development Southern Africa*, 37(6): 888-903.
- Nirmala, J. & Sengar, C. P. S. 2020. Label assignment and sentimental analysis for a product review on Twitter data. In: Springer. *Conference proceedings*. International Conference on Soft Computing and Signal Processing, Hyderabad, India. 199-208.
- Nortje, M. J. 2017. The effect of poverty on education in South Africa. *Educator Multidisciplinary Journal*, 1(1): 47-62.
- Nyakurukwa, K. & Seetharam, Y. 2023. Does online investor sentiment explain analyst recommendation changes? Evidence from an emerging market. *Managerial Finance*, 49(1): 184-204.
- Nyakurukwa, K. & Seetharam, Y. 2023. Investor reaction to ESG news sentiment: Evidence from South Africa. *Economia*, 24(1): 68-85.
- Nyasha, S. & Odhiambo, N. M. 2013. The dynamics of stock market development in the United States of America. *Risk governance & control: financial markets & institutions*, 3(1): 93-102.

- O'Donnell, F., Priestley, M. & Simperl, E. 2023. A survey of data quality requirements that matter in ml development pipelines. *ACM Journal of Data and Information Quality*, 15(2): 1-39.
- Obaida, T. H. & Salman, H. A. 2021. BBC News data classification using Naïve Bayes based on bag of words. *Journal of Human University (Natural Sciences)*, 48(9): 149-164.
- Ocran, M. K. 2010. South Africa and United States stock prices and the Rand/Dollar exchange rate. *South African Journal of Economic and Management Sciences*, 13(3): 362-375.
- Ojah, K. & Oluoch, W. 2024. Financial market development and the microstructure of corporate bond markets in Africa: A survey. *African Finance Journal*, 26(1): 1-15.
- Ortony, A. & Turner, T. J. 1990. What's basic about basic emotions? *Psychological Review*, 97(3): 315-331.
- Ortony, A. 2021. Are all "basic emotions" emotions? A problem for the (basic) emotions construct. *Perspectives on Psychological Science*, 17(2): 1-21.
- Ortony, A. 2022. Are all "basic emotions" emotions? A problem for the (basic) emotions construct. *Perspectives on Psychological Science*, 17(1): 41-61.
- Orts, M. Á. & Vargas-Sierra, C. 2023. Sentiment and emotion in financial journalism: A corpus-based, cross-linguistic analysis of the effects of COVID. *Humanities and Social Sciences Communications*, 10(219): 1-17.
- Othalasseril, D. & Shaikh, S. 2021. TradeZilla using algorithmic trading. In: Springer. *Conference proceedings. 2021 IEEE India Council International Subsections Conference (INDISCON)*, Nagpur, India. 1-6.
- Ozcelik, O., Şahinuç, F., Toraman, C. & Yilmaz, E. H. 2023. Impact of tokenization on language models: An analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4): 1-21.
- Pal, A. 2021. Logistic regression: A simple primer. *Cancer Research, Statistics, and Treatment*, 4(3): 551-554.
- Pamuji, A. 2021. Performance of the k-nearest neighbors method on analysis of social media sentiment. *JUISI*, 7(1): 32-37.
- Pan, Y. 2016. Heading toward Artificial Intelligence 2.0. *Engineering*, 2(4): 409-413.
- Parrott, G. W. 2001. Emotions in social psychology: Essential Readings. England, East Sussex.
- Parvathi, R. & Ramachandran, D. 2019. Analysis of Twitter specific pre-processing technique for Tweets. *Procedia Computer Science*, 165(10): 245-251.
- Patel, N. & Upadhyay, S. 2012. Study of various decision tree pruning methods with their empirical comparison in WEKA. *International Journal of Computer Applications*, 60(12): 20-25.

- Patel, R. B. 2023. From web scraping to machine learning: Approaches and tools for social network mining. *International Journal of Computer Applications*, 184(46): 35-44.
- Pawaskar, S. 2022. Stock price prediction using machine learning algorithms. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 10(1): 667-673.
- Phung, V. H. & Rhee, E. J. 2019. A high-accuracy model average ensemble of convolutional neural networks for classification of Cloud Image Patches on small datasets. *Applied Science*, 9(21): 1-16.
- Plutchik, R. 2001. The nature of emotions. *American Scientist*, 89(4): 344-350.
- Polak, K. 2021. The impact of investor sentiment on direction of stock price changes: Evidence from the Polish Stock Market. *Journal of Banking and Financial Economics*, 2(16): 72-90.
- Porter, M. E. 2008. The five competitive forces that shape strategy. *Harvard Business Review*. 86(1): 78-93.
- Poterba, J. M. 2000. Stock market wealth and consumption. *Journal of Economic Perspectives*, 14(2): 99-118.
- Prajapati, A. K. & Singh, U. K. 2023. An optimal solution to the overfitting and underfitting problem of healthcare machine learning models. *Journal of Systems Engineering and Information Technology (JOSEIT)*, 2(2): 77-84.
- Prats, M. A. & Sandoval, B. 2020. Does stock market capitalization cause GDP? A causality study for Central and Eastern European countries. *Economics: The Open-Access, Open-Assessment E-Journal*, 14(2020-17): 1-29.
- Prazak, T. 2020. The role of main microeconomic factors on the stock prices of selected Swiss companies. *Financial Internet Quarterly*, 16(1): 67-74.
- Purushotham, E., SB, A., Selvakumar, R. V., Srinivasan, A. & Shivaranjini. 2024. An evaluation of the CNN-LSTM Model's efficacy in sentiment analysis using the BERT and attendance mechanisms. *International Research Journal of Multidisciplinary Scope (IRJMS)*, 5(2): 822-829.
- Quadri, H. S. & Selvakumar, R. K. 2020. Performance of Naïve Bayes in sentiment analysis of user reviews online. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 10(2): 64-68.
- Raghvendra, K. 2024. Emotional intelligence and leadership effectiveness: Navigating decision-making, team dynamics, and organizational success. *International Journal of Science and Research (IJSR)*, 13(9): 1411-1416.
- Rahuljha. 2020. *LSTM gradients: Detailed mathematical derivation of gradients for LSTM cells*. <https://medium.com/data-science/lstm-gradients-b3996e6a0296> Date of access: 1 May 2025.

- Rajasekhar, K. 2022. Sentimental analysis on Instagram using machine learning techniques. *International Journal of Research Publication and Reviews*, 3(11): 72-976.
- Rathburn, P., Silberstein, S. & Wagner, H. 2022. *Why volatility is important for investors*. <https://www.investopedia.com/articles/financial-theory/08/volatility.asp> Date of access: 31 August 2022.
- Reddy, B. E., Reddy, P & Viswanath, P. 2018. Semi-supervised learning: A brief review. *International Journal of Engineering & Technology*, 7(1-8): 81-85.
- Reddy, S., Reddy, S., Shaik, S. & Sri, R. 2021. Sentimental analysis using logistic regression. *International Journal of Engineering Research and Applications*, 11(7): 36-40.
- Rösch, D. 2021. The impact of arbitrage on market liquidity. *Journal of Financial Economics*, 142(1): 195-213.
- Sacramento, J. F. 2021. *Sentiment analysis in the stock market based on Twitter data*. Lisboa: University Institute of Lisbon. (Thesis – MSc).
- Sadeghzadeh, K. 2018. The effects of microeconomic factors on the stock market: A panel for the stock exchange in Istanbul ARDL analysis. *Theoretical and Applied Economics*, 25(3): 113-134.
- Sathyanarayanan, S. & Tantri, B. R. 2024. Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, 27(4s): 4023-4031.
- Setiawan, E. B., Sina, M. N. I. 2023. Stock price correlation analysis with Twitter sentiment analysis using The CNN-LSTM method. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 7(4): 2190-2202.
- Sgammini, R. 2023. A comparative risk-adjusted performance evaluation of South African SRI funds and the FTSE/JSE over the Covid-19 Period. *International Journal of Economics and Financial Issues*, 13(1): 46-55.
- Shabrina, Z. & Qi, Y. 2023. Sentiment analysis using Twitter data: A comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, 13(1): 1-14.
- Shakhovska, K., Shakhovska, N. & Veselý, P. 2020. The sentiment analysis model of services providers' feedback. *Electronics*, 9(1922): 1-15.
- Shapira, I. 2023. Expressivity of shallow and deep neural networks for polynomial approximation. *ArXiv*, abs/2303.03544(1): 1-14.
- Sharma, N. & Tyagi, A. 2018. Sentiment analysis using logistic regression and effective word score heuristic. *International Journal of Engineering & Technology*, 7(2.24): 20-23.

- Shaziya, H. & Zaheer, R. 2021. Impact of hyperparameters on model development in deep learning. In: Springer. *Conference proceedings*. Proceedings of International Conference on Computational Intelligence and Data Engineering, Singapore, Asia. 57-67.
- Shinde, G. K. 2021. Sentiment analysis using hybrid approach. *International Journal for Research in Applied Science and Engineering Technology*, 9(12): 282-285.
- Singpurwala, K., 2021. *Sentiment analysis trading indicators*, Enschede: University of Twente.
- Sjong, D. 2021. Access denied? Unauthorized access after hiQ Labs v. LinkedIn. *Marquette Intellectual Property & Innovation Law Review*, 25(2): 133-147.
- Smarandache, F. 2024. Neutrosophy means: Common parts to uncommon things and uncommon parts to common things uncommon parts to common thing. *Neutrosophic Sets and Systems*, 68(1): 1-7.
- Solanki, M. S. 2019. Sentiment analysis of text using rule based and natural language toolkit. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(12S): 164-168.
- Speckmann, F. 2021. Web scraping: A useful tool to broaden and extend psychological research. *Zeitschrift für Psychologie*, 229(4), 241-244.
- Stringham, E. 2002. The emergence of the London Stock Exchange as a self-policing club. *Journal of Private Enterprise*, 17(2): 1-16.
- Sudhir, P. & Suresh, V. 2021. Comparative study of various approaches, applications and classifiers for sentiment analysis. *Global Transitions Proceedings*, 2(2): 205-211.
- Sullivan, M. 2021. Abstract. In: Kern, J. & Sullivan, M., eds. *The digital transformation of logistics: Demystifying impacts of the Fourth Industrial Revolution*. Wiley. 169-184.
- Tabot, E.S. 2022. Fundamental analysis from value drivers: the case of banking, industrial, consumer goods, healthcare and tech firms listed on the Johannesburg Stock Exchange. *Academy of Accounting and Financial Studies Journal*, 26(4): 1-11.
- Taherdoost, H. 2024. Beyond supervised: The rise of self-supervised learning in autonomous systems. *Information*, 15(8): 1-18.
- Talaat, A. S. 2023. Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data*, 10(1): 1-18.
- Tangirala, S. 2020. Evaluating the impact of GINI index and Information Gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2): 612-619.
- Tarale, H. 2025. Web scraping using Python. *International Journal of Multidisciplinary Research*, 3(1): 1023-1033.

- Taye, M. M. 2023. Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. *Computation*, 11(3): 1-23.
- Thapaliya, B. 2022. News content sharing on social media; Distribution or demonstration? A comparative study of Nepali and Norwegian newspapers. *Global perspectives on journalism in Nepal*, 252-268.
- Tiwari, S., Tiwari, T. & Tiwari, T. 2018. How Artificial Intelligence, machine learning and deep learning are radically different? *International Journals of Advanced Research in Computer Science and Software Engineering*, 8(2): 1-9.
- Tu, R., Yu, Q. & Zhan, M. 2018. Understanding readers: Conducting sentiment analysis of instagram captions. In: Association for Computing Machinery. *Conference proceedings*. 2018 2nd International Conference on Computer Science and Artificial Intelligence, Shenzhen, China. 33-40.
- Ukwandu, D. C. 2019. South Africa as a developmental state: Is it a viable idea? *African Journal of Public Affairs*, 11(2): 41-62.
- Van Den Broek, E. L. & Van Der Sluis, F. 2025. Model interpretability enhances domain generalization in the case of textual complexity modelling. *Patterns*, 6(2): 1-19.
- Van Rensburg, P. 2002. Market segmentation on the Johannesburg Stock Exchange II. *Studies in Economics and Econometrics*, 26(1): 83-100.
- Vergne, J. 2008. The social construction of competitive advantage and the origin of global capitalism: The case of the Dutch East India Company. *Academy of Management Proceedings*, 1(1-6): 1-8.
- Verhoef, G. 2012. Mutuality and regulation: The transition from mutual to public in the South African long-term insurance industry. *Journal of Economic and Financial Sciences*, 5(2): 567-590.
- Wang, K. 2022. Quantitative evaluation model of stock market liquidity by macroeconomic factors. *Mathematical Problems in Engineering*, 2022(6): 1-12.
- Wang, K. 2024. Empirical analysis of factors affecting stock liquidity-taking ping an bank as an example. *Advances in Economics Management and Political Sciences*, 79(1) :322-326.
- Wang, Z. 2023. Investor sentiment and the stock market. *Highlights in Business Economics and Management*, 21(1): 346-351.
- Wierzbicka, A. 2010. On emotions and on definitions: A response to Izard. *Emotion Review*, 4(2); 379-380.
- Wu, P. 2022. A survey of few-shot learning research based on deep neural network. *Frontiers in Computing and Intelligent Systems*, 2(1): 110-115.

- Xu, K., Zhao, J. & Zhou, Z. 2018. Tales of emotion and stock in China: Volatility, causality and prediction. *World Wide Web*, 21(3): 1093–1116.
- Xue, L. 2023. The analysis of relationship between emotional factors and stock price. *BCP Business & Management*, 38(1): 844-849.
- Yadav, A. 2023. Sentiment analysis using Twitter data. *International Journal For Science Technology And Engineering*, 11(5): 5833-5837.
- Yahoo Finance. 2025. *Yahoo Finance*. <https://finance.yahoo.com/> Date of access: 20 June 2024.
- Yu, Q. 2023. Progress of the study on the impact of investor sentiment on stock returns. *BCP Business & Management*, 39(1): 106-111.
- Yüksel, I. 2012. Developing a multi-criteria decision-making model for PESTEL analysis. *International Journal of Business and Management*. 24(7): 52-66.
- Zamora, A. 2019. Making room for big data: Web scraping and an affirmative right to access publicly available information online. *The Journal of Business, Entrepreneurship & the Law*, 12(1): 204-228.
- Zhao, B. 2017. Web Scraping. *Encyclopedia of big data*. 1-3.

APPENDIX A

This is to declare that I, Elma van Wyk, translator, language editor and interpreter, member no. 1002646 of the South African Translators' Institute (SATI),
have language edited the
Thesis submitted in fulfilment of the requirements for the degree Doctor of Philosophy in
Computer and Information Sciences with Computer Science and Information Systems
At the North-West University by

Jana Steenkamp

NWU number 26969513

with the title

**A machine learning approach to modelling social media sentiment for
stock market trading in South Africa**

Signature:



Date: 2025/11/24

Elma van Wyk (Ms), SATI-member no. 1002646

Tel. 082 952 1003