IST-Africa 2024 Conference Proceedings
Miriam Cunningham and Paul Cunningham (Eds)
IST-Africa Institute and IIMC, 2024
ISBN: 978-1-905824-73-1

Building a Dataset for Misinformation Detection in the Low-Resource Language

Mulweli MUKWEVHO, Seani RANANGA, Mahlatse S MBOOI, Bassey ISONG, Vukosi MARIVATE
Data Science for Social Impact, Department of Computer Science, University of Pretoria
Next generation enterprise and institutions, Council for Scientific and Industrial Research, Pretoria, South Africa
Department of Computer Science, North-West University, South Africa
Email: u23877716@tuks.co.za, seani.rananga@up.ac.za, vukosi.marivate@cs.up.ac.za, mratsoma@csir.co.za, Bassey.Isong@nwu.ac.za

Abstract: In the modern digital age, the widespread dissemination of misinformation has become a serious issue. Most focus in identifying misinformation online has been targeted at the English language in contrast to low-resource languages like Tshivenda. In this paper, we create a new dataset for news in the Tshivenda language to assist in developing resources for misinformation in the language. In our proposed methodology, we leveraged conditional random fields (CRF), gated recurrent unit (GRU), and long short-term memory (LSTM) to collect and annotate social media content. By applying these deep learning approaches to existing Tshivenda posts, we can assess their effectiveness for identifying false news in a low-resource language setting. This paper emphasises the vital need to combat misinformation in languages with limited resources, such as Tshivenda. Through the creation of a specialised dataset and the use of advanced techniques, it aims to address the problem of the spread of misinformation in low represented language communities.

Keywords: Misinformation, Natural Language Processing (NLP), social media, low-resource language, Conditional Random Fields (CRF), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM)

1. Introduction

Misinformation refers to the dissemination of inaccurate information with the intent to deceive [1]. This misinformation can be spread through a variety of channels, including social media, news websites, blogs, and other online platforms. This research has several important implications, including promoting critical thinking, protecting public health and safety, promoting informed decision-making, and strengthening democracy. In addition, the ease with which social media platforms allow users to share their opinions without verifying the accuracy of their ideas is significant [2].

Misinformation can spread rapidly and influence public opinion and even political decisions. Our research focus is on identifying misinformation in Tshivenda, a low-resource language [2]. The motivation for crawling a data set to detect misinformation in the Tshivenda language lies in the unique linguistic challenge presented by this language. In Tshivenda, one word has different meanings, leading to potential misunderstandings and hindering the effective sharing of intended content within the Tshivenda case study. This language often lacks proper fact-checking tools, making it more vulnerable to misinformation [3].

The choice of Tshivenda is particularly important in the South African context because it offers cultural diversity and has the potential to promote inclusive access to

information, thereby addressing linguistic inequalities in a multilingual society [4]. Empowering marginalised communities through the study of low-resource languages helps ensure the accuracy and equity of information and promotes global information justice. Figure 1 shows words that have different meanings which contribute to misleading and misunderstanding of the content in the Tshivenda language [35].

pfa … hear, feel, taste, understand; spit
vhafuwi … farmers; Chief
tama … wish, admire, desire, be eager, envy, prefer, crave

Figure 1: Words having different meanings in Tshivenda language [5][6].

This paper contributes to filling this gap by evaluating machine learning models for detecting misinformation in Tshivenda through web crawling. In addition, scientific studies have revealed the alarming frequency of misinformation related to COVID-19 on social media platforms where in South Africa, especially in rural and low-income areas [7], the spread of misinformation led some people to refuse vaccinations, claiming that the vaccines were unsafe or contained harmful ingredients [8].

The consequences of misinformation are not limited to the digital realm. For example, the xenophobic attacks in Johannesburg in 2019 were fueled in part by false news spread on social media [9]. Some of these news falsely attributed South Africa's high crime rate to foreigners, causing public resentment and anger. The severe riots and looting in July 2021 in KwaZulu Natal and Gauteng were also linked to the spread of false news on social media. misinformation on social media suggested that former President Jacob Zuma was going to be released from prison and this encouraged his supporters to take street protests [10].

The main goal of this paper was to propose advances in deep learning models, including CRF, GRU, and LSTM models. In the context of misinformation detection in the low-resource language Tshivenda, three different models CNN, GRU and LSTM are used in the work, each with unique strengths. Convolutional neural networks (CNNs) excel at capturing spatial patterns, enabling them to effectively recognize visual cues and patterns in social media content. Gated Recurrent Units (GRUs) focus on capturing sequential dependencies, allowing them to understand the temporal context of textual information, which is crucial given the dynamic nature of discourse in social media. Long Short-Term Memory (LSTM) networks, which are known for their memory performance, are able to capture long-term dependencies, which enables a sophisticated understanding of complex language structures [30].

By integrating these different models, a comprehensive approach that utilises spatial, temporal and contextual aspects is taken to improve the effectiveness of misinformation detection in the unique linguistic landscape of Tshivenda. These models assisted a lot in getting a crawled data set for fake information detected in low-resource languages, focusing on the Tshivenda case study. Previous research has shown that improvements in Natural Language Processing (NLP) mostly benefit high-resource languages, while low-resource languages such as Tshivenda have limited resources to detect misinformation [11] [12]. This gap poses a significant threat to society, leading to political instability, social unrest, and potential harm. Therefore, there is an urgent need to research practical methods for detecting misinformation in low-resource languages, with a particular focus on Tshivenda.

The following research questions were addressed:

(1) What are the challenges and methods for crawling a reliable dataset:
The project explores the challenges and effective methods associated with crawling a

reliable dataset for detecting misinformation in low-resource languages, with a focus on the Tshivenda language. This step is critical to understanding the particular obstacles that must be overcome in obtaining a high-quality data set [13].

(2) How is the data set being used to improve misinformation detection:
Once the data set is available, the project explores how it is used effectively to improve the accuracy of misinformation detection in low-resource languages. This includes developing, and implementing techniques and models such as LSTM, CRF, and GRU to effectively detect false news in the Tshivenda language [14].

2. Study Objectives

The aim of this research paper is to address the escalating problem of misinformation in resource-poor languages, with a focus on Tshivenda. The main goal is to combat the spread of misinformation for under represented languages in communities through creating a specialised dataset and using advanced deep learning techniques. This research aims to contribute to the broader discourse on the detection of misinformation in linguistically diverse contexts [29].

The study attempts to develop a well-annotated Tshivenda message dataset and implement Conditional Random Fields (CRF), Gated Recurrent Unit (GRU) and Long Short Memory (LSTM) models to evaluate the effectiveness of these deep learning approaches in detecting misinformation. By disseminating the results and findings, practical recommendations for combating misinformation in resource-constrained languages will be provided, emphasising the urgency of this task for preserving the integrity of information in Tshivenda-speaking communities [31][33].

In continuation of the research objectives, a structured approach is pursued to establish a specialised dataset for Tshivenda messages.
(1) Utilise advanced deep learning techniques, including CRF, GRU, and LSTM models.
(2) Evaluate the effectiveness of deep learning approaches in detecting misinformation.
(3) Disseminate research findings to contribute to the broader discourse on misinformation detection.
(4) Provide practical recommendations for combating misinformation in resource-limited languages
(5) Emphasise the urgency of addressing misinformation for preserving information integrity in Tshivenda-speaking communities

3. Methodology

In our approach, we began by crawling a dataset of Tshivenda posts from various social media platforms. This involved the collection of data relevant to our Tshivenda fake information detection case study. To ensure the dataset's quality, we removed irrelevant information such as stop words [15]. Figure 2 shows how web crawling was done.

In exploring fake information detection in low-resource languages, particularly in the Tshivenda language, we utilised the power of the Python integrated development environment (IDE) PyCharm [14]. This platform served as the hub for conducting all the experiments that were critical to our study. To ensure the efficacy of our research, we leveraged a suite of essential Python libraries, including but not limited to pandas, numpy, TensorFlow, sequential, dense, CRU, CRF, and LSTM, each playing a pivotal role in our model's development and evaluation [17].

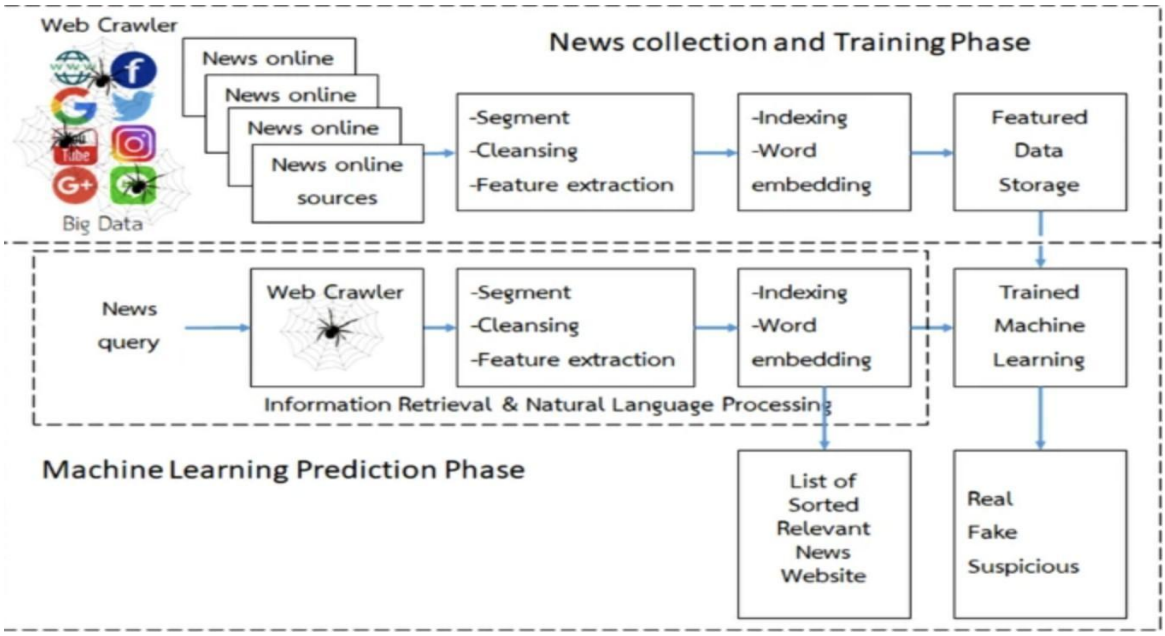

Figure 2: Web Crawling process [16].

Given the challenge of the limited availability of publicly accessible data sets in South African languages for researchers, we were compelled to source a dataset from an undisclosed origin, where web scraping and web crawling methods were used to extract misinformation from the Tshivenda case study. The following table shows the data set that was obtained after crawling from a different source [18].

Table 1: Obtained datasets after crawling process.

Name	**Description**
Limpopo Mirror	Discuss the farmers' strike against the municipality in Tshivenda, highlighting the unfortunate situation where community members, unaware of the strike, were unjustly arrested. ("Vhalimi vha Tshilamba vho sinyuswa nga uwela ha kholomo mulindini wa sworedzhi".)
Infonews	Web crawling of the data set from social media using CRF, GRU, AND LSTM.
Vuk'uzenzele	Explore the significance of combatting the COVID-19 virus within the Tshivenda language context. ("Kha vha thuse u fhelisa u phadalala ha COVID-19").
News.csv Github	A wide-ranging collection of news articles that cover events related to COVID-19, the 2016 United States presidential election, looting incidents during the arrest of the former president and their spread through social media.

For feature engineering, we focused on three specific models: CRF, GRU, and LSTM. These models were chosen because they excel at capturing contextual information, sequential dependencies, and long-term structural patterns in the data [19][34]. This selection was informed by a comprehensive literature review that we conducted to understand the state of the art in misinformation detection, with a specific focus on Tshivenda [3].

4. Results

We evaluated the performance of our models by comparing the outcome results of the authenticity of news articles to their actual labels (real or fake), where the percentages of the news that were declared real and fake were obtained during the training and training of the models. This evaluation process is instrumental in determining the accuracy and effectiveness of our approach. GRU tends to be the best in accuracy detection after training as is well-suited for handling sequences of data. The following tables show the different performances of LSTM, GRU, and CRF during training, and also after training of the model accuracy comparison which clearly shows which model is best in the detection of misinformation [23].

Table 2: LSTM Performance Metrics.

LSTM Performance	**Percentages(%)**
Classifier Accuracy	92.42
The news article is classified as 'Real'	92.42
The news article is classified as 'Fake'	7.58
Validation Accuracy	85.5
Training Accuracy	92.30
Accuracy during Training	88.70

Table 3: GRU Performance Metrics.

GRU Performance	**Percentages(%)**
Classifier Accuracy	92.58
The news article is classified as 'Real'	92.58
The news article is classified as 'Fake'	7.18
Validation Accuracy	85.50
Training Accuracy	91.80
Accuracy during Training	89.20

Table 4: CRF Performance Metrics.

CRF Performance	**Percentages(%)**
Classifier Accuracy	92.58
The news article is classified as 'Real'	92.58
The news article is classified as 'Fake'	7.18
Validation Accuracy	85.50
Training Accuracy	91.80
Accuracy during Training	89.20

Table 5: Overall performances

Models	Performance during training	Performance after training	Overall performance
LSMT	85%	92.42%	92.00%
CRF	89%	92.58%	92.74%
GRU	80%	92.82%	92.82%

In misinformation detection, text data is highly sequential, where the order of words and phrases matters. GRU's ability to capture long-term dependencies and contextual information in the text is crucial for making accurate predictions. It can learn and remember patterns and relationships in text, which is essential for discerning misinformation [20][28].

We chose the GRU model where the data set was taken from a social media platform. The following table consists of attributes and their description:

ATTRIBUTES	DISCRIPTION
ID	unique id for a news article
Title	the title Ofia news article
Author	author Ofithe news articl
Text	the information Ofinews article

Table 6: Data set attributes and their description obtained from social media [20].

Together, these attributes form a structured data set that enables the use of recognition methods such as CNN, GRU and LSTM. The unique ID helps to track and reference articles, while the title and author provide contextual information about the source. The text attribute as the core content is crucial for training and evaluating the misinformation detection models in the Tshivenda language. This organised social media dataset serves as the basis for robust analysis and effective misinformation detection in low-resource language.

The dataset consists of a total of 23 news articles for training and testing the model formed with the combination of real and misinformation [21]. CRF and LSTM models were employed for making predictions based on the input news, as these models were adapted for capturing sequential dependencies and contextual information for handling sequential data, capturing long-term dependencies, and understanding context and CRF models are strong for sequence labelling tasks [24][36].

In the final stages of our methodology, we combine web crawling, preprocessing, feature engineering, and deep learning techniques to create a comprehensive dataset tailored for fake information detection in Tshivenda based on the models being used which are GRU, CRF, and LSTM [27]. This dataset serves as a valuable resource, addressing the pressing challenge of misinformation in low-resource languages. By following this logical and systematic flow, our methodology is designed to contribute to the understanding and mitigation of misinformation in linguistic contexts often underrepresented in research and technology development. The following diagram shows how detecting real and misinformation was obtained using LSTM [32].

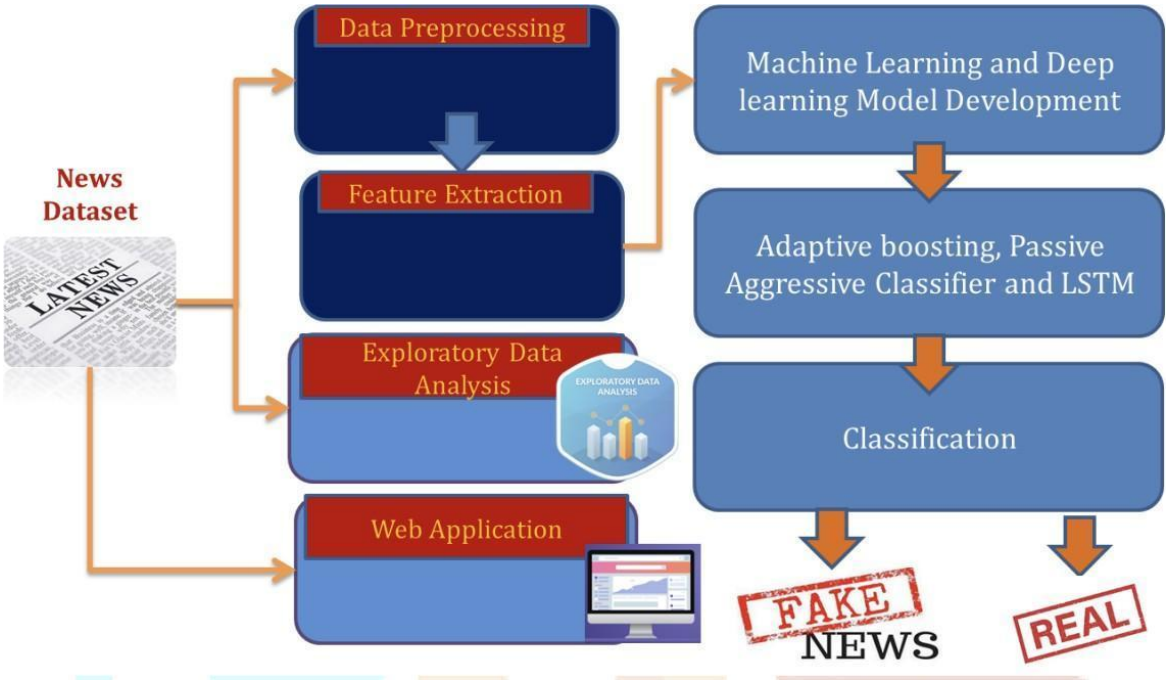

Figure 3: Detecting misinformation using LSTM [16][22].

5. Conclusions

Our exploration of crawling a dataset for fake information detection in the low-resource language of Tshivenda not only addresses the pressing need to combat misinformation but also successfully tackles the linguistic challenges unique to resource-poor languages [25]. Our research has strategically filled a critical gap by focusing on practical methods for detecting misinformation in a low-resource language, with a particular emphasis on Tshivenda [36]. Previous studies have revealed that advancements in Natural Language Processing (NLP) predominantly benefit high-resource languages, leaving low-resource languages like Tshivenda with limited resources for misinformation detection. This substantial gap has posed a considerable threat to society, manifesting as political instability, social unrest, and potential harm. Therefore, our research has taken a significant step in solving this issue by empowering marginalised communities and ensuring that they have the necessary tools to combat misinformation effectively. Our findings have implications not only for Tshivenda but for all resource-poor languages, safeguarding them from the pervasive influence of deceptive content [37][33]. As we navigate the digital landscape, it is our collective responsibility to ensure that all languages, regardless of their resource status, are fortified with the means to distinguish between authentic information and deceptive content, thus strengthening our collective defences against this persistent challenge [26].

Declaration of Use of Content generated by Artificial Intelligence (AI) (including but not limited to Generative-AI) in the paper

The authors acknowledge the use of content generated by Artificial Intelligence (AI) in the paper entitled "***Building a Dataset for Misinformation Detection in the Low-Resource Language***" in the following ways:

- Text Generation and Editing: ChatGPT was used to generate initial drafts of certain sections of the paper, including the background and literature review. These sections were subsequently reviewed and edited by the authors to ensure accuracy and coherence.
- Grammar and Style Enhancement: Grammarly was used to check for grammatical

errors and improve sentence structure, Instant-text to paraphrase and style the text.

- Data Visualization: DataRobot was used to create and optimize figures and visual representations of data.

References

[1] E. M. Savino, "misinformation: No one is liable, and that is a problem," Buff. L. Rev., vol. 65, p. 1101, 2017.

[2] P. Meel and D. K. Vishwakarma, "misinformation, rumor, information pollution in social media and web: A contemporary survey of state-of-the arts, challenges and opportunities," Expert Systems with Applications, vol. 153, p. 112986, 2020.

[3] E. J. Malatji, The impact of social media in conserving African languages amongst youth in Limpopo Province. PhD thesis, 2019.

[4] M. D. Patterson, " It's that they participate intellectually mostly in Tshivenda": indigenous multilingual education in Vhembe, South Africa: a thesis submitted in partial fulfilment of the requirements for the degree 18 of Master of International Development, Department of People, Environment & Planning at Te Kunenga Ki Purehuroa–Massey University, Aotearoa New Zealand. PhD thesis, Massey University, 2021.

[5] M. T. Makhavhu, Relational nouns in Tshivenda. PhD thesis, Stellenbosch: Stellenbosch University, 2006.

[6] M. Madiba and D. Nkomo, "The tshivenda–english thalusamaipfi/dictionary as a product of south african lexicographic processes," Lexikos, vol. 20, pp. 307–325, 2010.

[7] T. M. Silubonde, L. Knight, S. A. Norris, A. van Heerden, S. Goldstein, and C. E. Draper, "Perceptions of the covid-19 pandemic: a qualitative study with south african adults," BMC Public Health, vol. 23, no. 1, pp. 1–15, 2023.

[8] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual sentiment analysis for under-resourced languages: A systematic review of the landscape," IEEE Access, 2022.

[9] V. Chenzi, "Misinformation, social media and xenophobia in South Africa," African Identities, vol. 19, no. 4, pp. 502–521, 2021.

[10] M. T. Bhuda, T. Motswaledi, and P. Marumo, "Moral decay, government, and looting in south africa during covid-19," African Journal of Development Studies (formerly AFRICA Journal of Politics, Economics and Society), vol. 2023, no. si2, pp. 57–74, 2023.

[11] M. A. Hedderich, L. Lange, H. Adel, J. Strotgen, and D. Klakow, "A survey on recent approaches for natural language processing in low resource scenarios," arXiv preprint arXiv:2010.12309, 2020.

[12] V. Marivate, M. Mots' Oehli, V. Wagner, R. Lastrucci, and I. Dzingirai, "Puoberta: Training and evaluation of a curated language model for setswana," arXiv preprint arXiv:2310.09141, 2023.

[13] F. Gereme, W. Zhu, T. Ayall, and D. Alemu, "Combating misinformation in "low-resource" languages: Amharic misinformation detection accompanied by resource crafting," Information, vol. 12, no. 1, p. 20, 2021.

[14] J. E. Daniel, Applications of natural language processing for low-resource languages in the healthcare domain. PhD thesis, Stellenbosch: Stellenbosch University., 2020.

[15] K. Ghosh and A. Bhattacharya, "Stopword removal: Why bother? a case study on verbose queries," in Proceedings of the 10th Annual ACM India Compute Conference, pp. 99–102, 2017.

[16] P. Meesad, "Thai misinformation detection based on information retrieval, natural language processing and machine learning," SN Computer Science, vol. 2, no. 6, p. 425, 2021.

[17] D. H. NOVA VALCARCEL, "Anomaly detection system for automotive can using lstm autoencoders," 2019.

[18] R. Gray, R. Kouhy, and S. Lavers, "Constructing a research database of social and environmental reporting by uk companies," Accounting, Auditing & Accountability Journal, vol. 8, no. 2, pp. 78–101, 1995.

[19] J. Tang, F. Belletti, S. Jain, M. Chen, A. Beutel, C. Xu, and E. H. Chi, "Towards neural mixture recommender for long range dependent user sequences," in The World Wide Web Conference, pp. 1782–1793, 2019.

[20] A. Choudhary and A. Arora, "Linguistic feature based learning model for misinformation detection and classification," Expert Systems with Applications, vol. 169, p. 114171, 2021.

[21] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and misinformation using text classification," Security and Privacy, vol. 1, no. 1, p. e9, 2018.

[22] P. Bahad, P. Saxena, and R. Kamal, "misinformation detection using bidirectional lstm-recurrent neural network," Procedia Computer Science, vol. 165, pp. 74–82, 2019.

[23] Sangiri, Ratna, and Netai Mandal. "International journal of web information systems: An evaluation." KIIT Journal of Library and Information Management 9.1 (2022): 27-36.

[24] Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. "A Hybrid Model for Effective Fake News

Detection with a Novel COVID-19 Dataset." ICAART (2). 2021.
[25] Girgis, Sherry, Eslam Amer, and Mahmoud Gadallah. "Deep learning algorithms for detecting fake news in online text." 2018 13th international conference on computer engineering and systems (ICCES). IEEE, 2018..
[26] Strassel, Stephanie, and Jennifer Tracey. "LORELEI language packs: Data, tools, and resources for technology development in low resource languages." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016.
[27] Magueresse, Alexandre, Vincent Carles, and Evan Heetderks. "Low-resource languages: A review of past work and future challenges." arXiv preprint arXiv:2006.07264 (2020).
[28] Al-Masri E, Mahmoud QH. WSCE: A crawler engine for large-scale discovery of web services. InIEEE International Conference on Web Services (ICWS 2007) 2007 Jul 9 (pp. 1104-1111). IEEE.
[29] Levene, Mark, and Alexandra Poulovassilis. "Web Dynamics [electronic resource]: Adapting to Change in Content, Size, Topology and Use."
[30] Garrabé, É., & Russo, G. (2023). CRAWLING: a crowdsourcing algorithm on wheels for smart parking. Scientific Reports, 13(1), 16617.
[31] Mesham S, Hayward L, Shapiro J, Buys J. Low-resource language modelling of south african languages. arXiv preprint arXiv:2104.00772. 2021 Apr 1.
[32] Aïmeur, Esma, Sabrine Amri, and Gilles Brassard. "Fake news, disinformation and misinformation in social media: a review." Social Network Analysis and Mining 13, no. 1 (2023): 30.
[33] Di Domenico, Giandomenico, Jason Sit, Alessio Ishizaka, and Daniel Nunan. "Fake news, social media and marketing: A systematic review." Journal of Business Research 124 (2021): 329-341.
[34] Sides, John, Michael Tesler, and Lynn Vavreck. "The 2016 US election: How Trump lost and won." Journal of Democracy 28, no. 2 (2017): 34-44..
[35] Barve Y, Saini JR. Healthcare misinformation detection and fact-checking: a novel approach. International Journal of Advanced Computer Science and Applications. 2021;12(10).
[36] Bouvier G. What is a discourse approach to Twitter, Facebook, YouTube and other social media: connecting with other academic fields?. Journal of Multicultural Discourses. 2015 May 4;10(2):149-62.
[37] E. P. Lahiff, Agriculture and rural livelihoods in a South African Homeland': A case study from Venda. University of London, School of Oriental and African Studies (United Kingdom), 1997.