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

Generalising across Domains in Video Misinformation Detection

David WALKER, Seani RANANGA, Bassey ISONG, Vukosi MARIVATE
Data Science for Social Impact, Department of Computer Science,
University of Pretoria, North West University, South Africa
Email: u19055252@tuks.co.za, seani.rananga@up.ac.za, Bassey.Isong@nwu.ac.zav,
vukosi.marivate@up.ac.za

Abstract: This paper proposes and demonstrates the application of pre training and transfer learning in a deep neural network model to identify misinformation in YouTube videos, based on their auto-generated captions. Pre-training, in this context, is the training of smaller sub-models on domain-specific data, before those smaller models, together with all trained weights, are integrated together into one larger model. Transfer learning takes place when the larger model is able to utilise the knowledge gained by sub-models to generate new conclusions. The given model also utilises pre-trained Bidirectional Encoder Representations from Transformers (BERT) layers, further utilising the idea of pre-training and leveraging the concept of transfer learning to maximise efficacy. An appreciable improvement in overall quality is demonstrated when the proposed method is used, as compared to a naive model built to tackle this same problem. The goal of this research is to better understand the extent to which the given techniques and model architectures contribute to effective classification of misinformation across a variety of different domains.

Keywords: misinformation detection, video, neural network, deep learning, generalisation, YouTube, captions.

1. Introduction

The problem of generalising a misinformation classifier to be able to work in multiple domains of knowledge is an important one, due to the wide variety of misinformation which can be spread over the internet. However, effective generalisation has been difficult to achieve, with past examples of fake news detection models not generalising across multiple domains well [1]. This paper proposes and tests two methods of improving neural net misinformation classification, specifically in the domain of video misinformation, whereby captioning is used to detect said misinformation. It shows that pre-training a model based on known data about a domain can improve the model's net domain understanding. It also shows that combining and leveraging the individual domain knowledge of smaller models can lead to a larger model being more capable of generalisation overall. Ultimately, using the concepts discussed in this paper allows for better model scalability when new misinformation domains are discovered.

It has been previously established that fake news detection models tend not to generalise particularly well, as demonstrated by Hoy et al. [1]. Given that various examples of transfer learning [2] for misinformation detection have proven effective before [3], this paper proposes a model which uniquely trains small sub-models on specific domains of misinformation, before combining them to attempt a transfer learning method for detecting misinformation. This paper's primary objective is to determine whether such an approach is viable, and whether it leads to notably better results than a naive attempt to solve the problem without transfer learning.

Furthermore, this paper aims to answer whether utilising merely video captions for speech-dense videos allows for effective misinformation detection. It thus builds upon the dataset provided by Hussein et al. [4]. This model is largely inspired by the approach taken by Jagtap et al. [5] and leverages the word embeddings developed as part of the BERT project [6].

This model is built to be expanded: due to dataset limitations, only five domains of misinformation knowledge were trained, with those domains outlined in table 1. Each of those had a relatively small amount of training: thus, each individual stream can benefit from a larger dataset and more training. Furthermore, one of the main objectives of this approach is allowing for easier expansion. In that sense, new domains can have their own streams trained and added, and only the concatenation layers at the end will need to be retrained. This makes for faster and more scalable expansion of the model when new misinformation domains are discovered.

1.1 Related work

There are several existing models which tackle the issue of misinformation in video. Their characteristics are detailed in Table 1. Several datasets around video misinformation also exist and are described in Table 2.

Table I: Past Misinformation Detector Models In Videos.

Title	Year	Model used	Future work
Effective fake news video Detection using domain knowledge and multimodal data fusion on YouTube [7]	2022	Multi-layer perceptron	Use of subtitles and training to become domain agnostic.
A Deep Learning Framework for Detection of COVID-19 Fake News on Social Media Platforms [8]	2022	LTSM, bidirectional LTSM, CNN, and hybrid CNN/LTSM.	Using transformers as a replacement for LSTM models and including more languages and other countries in the training dataset.
A CNN based misleading video detection model [9]	2022	CNN	Training the model to work on platforms other than Billibilli.
AMultimodal Misinformation Detector for COVID-19 Short Videos on TikTok [10]	2021	Transformers, CNN, and LSTM fusion.	None explicitly mentioned.
FNDNet - a deep CNN for fake news detection [11]	2020	Deep CNN	Implementing relationship or network- gathering features into the model.
Misinformation Detection on YouTube Using Video Captions [5]	2021	Synthetic Minority Oversampling	Improvement of word embeddings.

Table II: Entries per Misinformation Domain in Dataset

Name	Description
InVID Fake Video Corpus 2018	A corpus of videos made to offer perspective on different types of fake video which may be found around the internet, containing 200 fake videos and 180 real ones. [12]
YouTubeAudit-data	A corpus of primarily metadata describing videos which were used in a paper by Hussein et al. [4] for misinformation detection.
FakeSV	A large corpus of Chinese short-form videos used by Qi et al. [13] for their own fake news detection implementation.
A Dataset of COVID- Related Misinformation Videos and their Spread on Social Media	A corpus containing metadata associated with COVID-related YouTube videos which were taken down due to their false information in their content. [14]

2. Objectives

In our research, we aim to propose a method for the accurate classification of misinformation in videos. This involves testing the effectiveness of utilising captions as a means to enhance the accuracy of misinformation detection. Additionally, we will explore the impact of pretraining on the model's performance. Our goal is to propose a model that not only generalises across various domains but also scales effectively to handle diverse datasets. To validate our approach, we will benchmark the proposed model against a baseline to demonstrate its efficacy and improvements in misinformation classification.

3. Methodology

3.1 Data Preparation

The decision was made to utilise the YouTube-Audit dataset [4]. It was chosen as it offered the ability to utilise the captions which had been automatically generated for each YouTube video, simplifying the feature extraction process while still providing useful information about any video which includes audible speech. Given that a video is by nature multimodal, any model which operates on it will need to use some decomposition of the video's features. This could take the form of visual features, audio features, metadata features or any other element of the video. In this case, given the past work done by Jagtap et al [5], the decision was made to utilise captions generated from the video's audio features. These captions are text-based and offer insight into any verbal communication which occurs during the video. This paper thus also provides an additional layer of research regarding the extent to which captions as a video feature can be used for misinformation classification. Initially, we built a new scraper to download the captions, utilising the datasets provided by Hussein et al [4]. However, pre-existing captions from Jagtag et al. were ultimately used [5]. The existing captions saved on processing power and helped overcome a dataset imbalance attributed to the takedown of many misinformation videos from YouTube. Hence, a total of 2672 entries were available. Table 3 describes how many entries were available for each given domain:

Table III: Entries per Misinformation Domain in Dataset

Misinformation domain	Available entries	Misinformation	Neutral	Debunking
9/11	436	51	336	49
Chemtrails	484	44	300	140
Flat earth theory	317	63	232	22
Moon landing	317	70	226	21
Vaccines	621	170	404	47

It is noted that there is an imbalance in the dataset. For our purposes, the dataset was not augmented, as to provide a realistic representation of the video proportion likely to be found on YouTube, and to examine whether the objective of generalisation by our proposed model can overcome this barrier more effectively than a naive model could.

However, augmentation could be utilised to create an even more effective model for real-world applications.

Within each domain, the dataset was split into testing and training datasets, with 90% of the data being assigned to the training dataset and 10% to the testing dataset. The testing data was never shown to the dataset directly: however, it did act as an early stopping mechanism, whereby the model was tested against it after each epoch. Should the error stagnate in respect to the testing set, training was halted for that sub-model.

The dataset features used were only the captions and the normalised annotation columns, as they contained all the necessary data for the implementation of the proposed

model. Further accuracy can likely be attained by including additional information from other parts of the dataset in the training data for this model.

4. Technology Description

This implementation was programmed in TensorFlow, running on Python. The implementation was trained for a maximum of 50 epochs, with early stopping criteria should the validation loss stagnate for 4 epochs. This includes each stream, which were pre-trained with the same maximum number of epochs and using the same stopping criteria as mentioned a moment ago. All networks were trained on an Nvidia GTX 1060 6GB.

Bidirectional Encoder Representations from Transformers (BERT) is used as a pre-trained encoder for text processing in this model. This takes inspiration from the various encoding methods used by Jagtag et al. [5] but changes the specifics regarding which encoder it is. BERT was chosen due to its accessibility, and because the bidirectional transformer architecture employed by the BERT project helps overcome some of the limitations of older pre-trained models. Specifically, the BERT encoding implementation is split into two layers, starting with a preprocessing layer, to strip caption data of unexpected characters and symbols and otherwise ensure that the string input is compatible with the encoding layer. The encoding layer itself is next, which encodes the strings. The sequence output of this layer is then used for the rest of the stream, representing each token in context, as opposed to the pooled output, which would represent each input sequence in its entirety.

Furthermore, for both the baseline and pretrained models, long short-term memory (LSTM) networks are used. This is due to past evidence proving the effectiveness of combining word embeddings with such networks for text classification tasks [16]. The use of memory cells in LSTM networks enables them to retain contextual information about whichever input they are processing. This leads to LSTM approaches being particularly effective for natural language processing (NLP) tasks [17]. This, in turn, should allow the network to gain a good contextual understanding of any given captions being tested, and thus better understand whether said captions are likely to be misinformation.

5. Developments

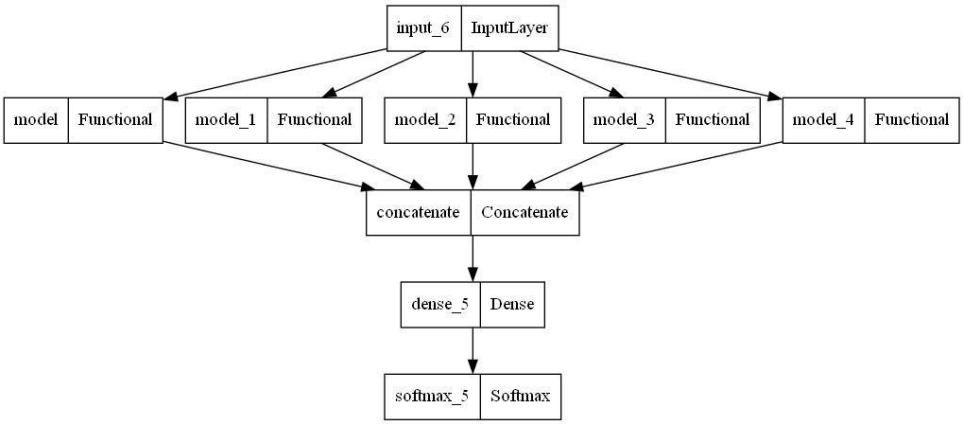

Figure 1-. An overview of the concatenation model

This model is intended to process generated text captions of videos. Thus, the first layer of this model is an input layer, which accepts said strings of variable length, and pads them appropriately. Thereafter, each domain splits into a separate "stream", each with its own trained sub-model. Each of these streams consist of a Bidirectional Encoder Representations from Transformers (BERT) based preprocessor and encoder, to encode the text features of the given input string. The stream then contains a dropout layer, in hopes of

reducing overfitting, followed by an LSTM layer [15], another dropout layer, a densely connected summation layer, and lastly a Softmax layer to determine which class (being misinformation, neutral, or debunking of misinformation) that submodel believes the input falls into. This stream is repeated in a parallel manner for each domain the network is meant to operate on. The final model then concatenates all these different models, again performs a summation in a densely connected layer, and then uses a final Softmax layer to determine the overall prediction of the model. Figure 1 provides a visualisation of the architecture.

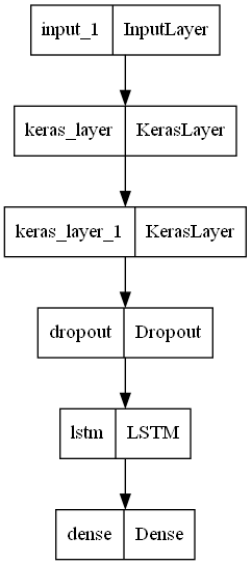

Figure 2 - An overview of an inner "stream" model

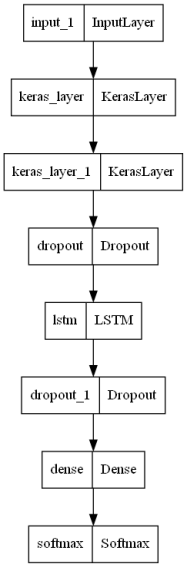

Figure 3- An overview of the comparison model

This model was developed with the intention of the sub-models or "streams" being tasked with the majority of misinformation classification work, as each individual stream is trained on a specific domain of misinformation and is meant to predict whether any input is misinformation in that domain. In their respective domains, these sub-models are effective. However, such an approach cannot generalise well: a network trained on one domain has little hope of understanding another. This is where the stream architecture becomes useful. Firstly, it allows the concatenation and densely connected layers which come after the streams to act as "combined likelihood estimators", whereby they can learn that the input is likely misinformation either if one stream is quite certain of that being the case, or potentially if multiple streams all have some indication of that possibility. (The same logic would apply to neutral and debunking classes.) Hence, an additional stream for a new domain of misinformation can easily be added to the model, should the ability or need to train one arise. Additionally, it allows the training process of this final layer to be less intensive, and the useful work which has gone into training past streams to be preserved. This minimises the computational cost of further generalising the model, while also leading to improved overall performance, as described in the results section.

The proposed model is compared to a linear LSTM implementation without multiple pre-trained streams, as described in Figure 3.

6. Results

6.1 Metrics [18]

Precision: Precision is the ratio of true positives (i.e., positive samples which have been classified correctly) to the total number of positives classified (i.e., true and false positives). It can be expressed as:

$$\frac{TP}{(TP + FP)}$$

Recall: Recall describes the proportion of "positives" which were identified. So, given *TP* as "true positives" (i.e., correctly classified entries) and *FN* as "false negatives" (i.e., entries which were meant to be in a class but were not classified as such), recall can be expressed as:

$$\frac{TP}{TP + FN}$$

F1-score: F1-score is the weighted average of precision and recall. More specifically, it is the weighted harmonic mean of the two measures. It accounts for both false positives and false negatives (through precision and recall), providing a balanced measure of a model's performance.

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

Weighted vs. macro average: The macro average is the average taken over all classes, whereby first the results are obtained for each class, and those results are then averaged. The weighted average, on the other hand, assigns a weight to the results of each class according to how many samples were tested, and in so doing, it provides proportionality to the result.

Table IV: Classification Report of Pretrained Model

	Precision	Recall	F1-score
Misinformation	0.43	0.48	0.46
Neutral	0.82	0.93	0.87
Debunking	0.67	0.13	0.22
Accuracy	0.75		
Macro average	0.64	0.51	0.51
Weighted average	0.74	0.75	0.71

This classification report shows promise for the proposed pre-training method, indicating a successfully trained model, which generalises across all five given domains of misinformation effectively.

Table V: Classification Report of Naive Model

	Precision	Recall	F1-score
Misinformation	0.32	0.69	0.44
Neutral	0.89	0.84	0.86
Debunking	0	0	0
Accuracy	0.72		
Macro average	0.40	0.51	0.43
Weighted average	0.7	0.72	0.70

Relative to the pre-trained model, with a similar setup, the naive model underperforms. Notably, this is not a commentary on the naive model being incapable: more extensive training, and a larger architecture, may all help the naive model to be more effective. However, as a baseline comparison, this "simpler" model does not perform quite as well as the pretrained model, potentially due to the breadth of the domains of knowledge it is tasked with understanding. In demonstrating this outcome, the improvement which can be attained in the realm of video misinformation classification is quite clear.

7. Business Benefits

Given the ever-growing problem of misinformation detection, many online platforms are either choosing to or being forced to integrate misinformation checks to their systems. However, businesses need to optimise their systems for maximum efficiency, to obtain maximum utility from their hardware. A scalable and modifiable system is thus preferred to a "static" one with no modular elements. To this end, the proposed model achieves the dual purpose of reducing the amount of additional training time needed to integrate new or updated misinformation domains to such systems, creating modularity, while also improving effectiveness of such a system. Ultimately, this has the potential to both save on training time for frequently updated models, and to ensure that models are at their most effective.

8. Conclusions

This model has proven acceptably effective at generalising misinformation across multiple domains. By doing so, it gave merit to the idea that pre-training helps misinformation detection models to better understand their individualised domains, leading to a more rounded and trustworthy prediction, with a 5.7% weighted precision improvement in the given test case. On that basis, then, the multi-stream architecture has also shown its value. By utilising this architecture, and simply integrating additional streams, the ability for training time to be minimised in future expansion of a model has been achieved, improving scalability. The goal of further exploring the use of video captions for misinformation classification has also been achieved, as this paper demonstrates that the use of auto-generated YouTube captions allowed the model to achieve a weighted average precision of 0.74.

This architecture is limited in that it still requires a degree of specialisation for each domain it wishes to classify misinformation in. However, the difficulty thereof is reduced.

In the future, different types of misinformation can be tested against this architecture, to determine whether the generalisation capabilities apply to broader realms of misinformation. Comparisons can also be made against less naive implementations of existing classifiers. Furthermore, larger, and more comprehensive datasets, and even combinations of disparate datasets, may be an interesting expansion on this idea, to decide whether this generalisation works in more "real-world" environments.

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

The authors confirm that there has been no use of content generated by Artificial Intelligence (AI) (including but not limited to text, figures, images, and code) in the paper entitled "***Generalising across Domains in Video Misinformation Detection***".

References

[1] N. Hoy and T. Koulouri, "Exploring the generalisability of fake news detection models," in 2022 IEEE International Conference on Big Data (Big Data), pp. 5731–5740, IEEE, 2022.

[2] L. Torrey and J. Shavlik, "Transfer learning," in Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pp. 242–264, IGI global, 2010.

[3] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in Proceedings of the AAAI conference on artificial intelligence, vol. 34, pp. 13915–13916, 2020.

[4] E. Hussein, P. Juneja, and T. Mitra, "Measuring misinformation in video search platforms: An audit study on youtube," Proceedings of the ACM on Human-Computer Interaction, vol. 4, no. CSCW1, pp. 1–27, 2020.

[5] R. Jagtap, A. Kumar, R. Goel, S. Sharma, R. Sharma, and C. P. George, “Misinformation detection on youtube using video captions,” arXiv preprint arXiv:2107.00941, 2021.
[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
[7] H. Choi and Y. Ko, “Effective fake news video detection using domain knowledge and multimodal data fusion on youtube,” Pattern Recognition Letters, vol. 154, pp. 44–52, 2022.
[8] Y. Tashtoush, B. Alrababah, O. Darwish, M. Maabreh, and N. Alsaedi, “A deep learning framework for detection of covid-19 fake news on social media platforms,” Data, vol. 7, no. 5, p. 65, 2022.
[9] X. Li, X. Xiao, J. Li, C. Hu, J. Yao, and S. Li, “A cnn-based misleading video detection model,” Scientific Reports, vol. 12, no. 1, p. 6092, 2022.
[10] L. Shang, Z. Kou, Y. Zhang, and D. Wang, “A multimodal misinformation detector for covid-19 short videos on tiktok,” in 2021 IEEE International Conference on Big Data (Big Data), pp. 899–908, IEEE, 2021.
[11] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, “Fndnet–a deep convolutional neural network for fake news detection,” Cognitive Systems Research, vol. 61, pp. 32–44, 2020.
[12] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and I. Kompatsiaris, “A corpus of debunked and verified user-generated videos,” Online information review, vol. 43, no. 1, pp. 72–88, 2019.
[13] P. Qi, Y. Bu, J. Cao, W. Ji, R. Shui, J. Xiao, D. Wang, and T.-S. Chua, “Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms,” arXiv preprint arXiv:2211.10973, 2022.
[14] A. Knuutila, A. Herasimenko, H. Au, J. Bright, and P. N. Howard, “A dataset of covid-related misinformation videos and their spread on social media,” Journal of Open Humanities Data, vol. 7, 2021.
[15] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[16] J.-H. Wang, T.-W. Liu, X. Luo, and L. Wang, “An lstm approach to short text sentiment classification with word embeddings,” in Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018), pp. 214–223, 2018.
[17] L. Yao and Y. Guan, “An improved lstm structure for natural language processing,” in 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), pp. 565–569, IEEE, 2018.
[18] H. Dalianis and H. Dalianis, “Evaluation metrics and evaluation,” Clinical Text Mining: secondary use of electronic patient records, pp. 45–53, 2018.