UNIVERSITY POLITEHNICA OF BUCHAREST
FACULTY OF AUTOMATIC CONTROL AND COMPUTERS
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

Scientific Report #1

Network-aware Fake News mitigation on social media
Artificial Intelligence

Octavian Mitrică

Scientific advisors:

S.L. Dr. Ing. Ciprian-Octavian Truică
Conf. Dr. Ing. Elena-Simona Apostol

BUCHAREST
2024

CONTENTS

1 INTRODUCTION

Fake news has always been a topic of interest. With social media platforms' continuous growth in popularity, the challenges posed by outlets spreading harmful content are becoming even more pronounced [1]. Fake news is defined as false or misleading information presented as genuine news and is used to exploit beliefs, biases, and emotional triggers to capture the audience's attention. In this research paper, we will explore different methods of detecting fake news, using different approaches, with a special focus on transformers [2] and the similarity between text and image in the context of social media posts.

The main objectives of this paper will be to experiment with creating different prediction models for:

- *1:* Text-only fake news detection
- *2:* Image-only fake news detection
- *3:* Multimodal fake news detection

This Scientific Report is structured in 6 chapters. Chapter 2 presents a survey and introduction into relevant work and State of the Art. Chapter 3 showcases the algorithms and models used for gathering predictions, as well as the datasets used for experiments. Chapter 4 describes the proposed solution and implementation of the models. Chapter 5 ranks the results of the model and explains the architecture and setup details. Finally, chapter 6 sums up all the gathered information, limitations, and further work.

2 STATE OF THE ART

In this chapter, we present the previous approaches and research work on the mentioned tasks of interest and their results, split into three categories of detecting fake news: text, image, and the combined analysis of text and image.

1. Text-only fake news detection

Text-only detection refers to the task of classifying a text input in the context of social media posts into *actual news* or *fake news*. Experiments around this area include text preprocessing and vectorization using word embedding techniques, NLP, and ML methods such as CNNs or LSTMs to predict an output. This task is most commonly documented and tested throughout research papers, with good resulting metrics all around.

In the article, [3], the authors present a new model for identifying fake news, using Natural Language Processing techniques, Graph Neural Networks, and Knowledge graphs. The variety of embedded encodings showed an improved performance in detection on the Kaggle Fake News dataset.

Truică et al. [4] propose a new deep learning model, which takes in word embeddings generated with Word2Vec pre-trained and feeds them to a Bidirectional Long Short-Term Memory (BiLSTM) layer. It may sound complicated but it is a layer with two simple LSTM architectures that can look both forward and backward in the sequence. Their model yields state-of-the-art results in detecting fake news on datasets such as Kaggle, GossipCop, and Fakeddit.

Article [5] proved enhanced results in fake news detection with transfer learning techniques and fine-tuning the transformer architectures for sentence embedding. Specifically, they experimented with BERT, XLNet, and RoBERTa transformers on the FNC-1 extended dataset.

2. Image-only fake news detection

Image-only detection refers to the task of classifying an image input in the context of social media into *actual news* or *fake news*. Common methods for handling this task are image preprocessing, visual feature extraction, and visual structure extraction, along with ML methods such as CNNs or transformers, for training and predicting an output. Detecting fake news based solely on image data is not such a popular topic, with text being more descriptive and telling.

Jin et al. [6] presented a new approach by extracting visual and statistical features to describe different patterns that show in images associated with fake news. Although their work uses predefined features and is not scalable for more complex datasets, it still proves relevant in the

task of detection.

Article [7] proposed a multi-domain visual neural network that combines frequency and pixel domains for detecting fake news with only visual characteristics. Their model automatically extracts the image quality features in the frequency domain and the image semantics in the pixel domain using CNNs.

3. Multimodal fake news detection

Since our focus is on the social media domain, news articles or posts usually come with both text and image data. To fuse textual and visual features for better performance, we need to first ask: are the text and image duo related in any way? The post can contain harmful content in either of the attributes so we need to establish their relationship. Recently, more research has emerged on such multimodal approaches and has proved the combination to be relevant.

Nakamura et al. [8] presented a novel multimodal dataset that consists of over 1 million samples of Reddit posts, spread into multiple categories. The samples have 2-way, 3-way, and 6-way classification labels for fake news. The authors also constructed multiple hybrid text and image models for classification. Their experiments include using InferSent and BERT to generate text embeddings for the title of the posts and VGG16, EfficientNet, and ResNet50 for extracting visual features. In the end, BERT for textual data and ResNet50 for visual data proved to be the best-performing ones out of the bunch on multimodal classification. We will also use their dataset later in the experiments section.

In the article [9], authors performed extensive experiments on the Fakeddit dataset mentioned above, involving text embedding and image features extraction, adding image captions and sentiment analysis in the mix. They finally designed a model that outperformed state-of-the-art models working on a similar dataset. It uses the Xception model to identify images with high digital alterations, BERT for learning contextual knowledge, and visual sentiment analysis to learn features that distinguish an image with negative readings.

Article [10] takes on the challenge of transformer fusion between text transformers and image transformers, expanding on the caption-based enhancement tactic. The authors looked into late and early fusion of the models and analyzed previous methods. Their design of the model consists of embedding the text data with BERT pre-trained and embedding the image data with ResNet and Faster RCNN to generate entities (Figure 1). With these entities and embeddings, they form a new embedding by combining them. A multi-modal transformer is then applied to the final embeddings for computing correlations between elements through a multi-head self-attention mechanism. They also experimented on the Fakeddit dataset and we will compare it with our results in a later section.

Since we are working with social media articles, we need to take a look at posts that contain more than one image as well. Anastasia et al. [11] introduced the idea of combining the visual information from multiple images. For image representation, they used a VGG16 pre-trained

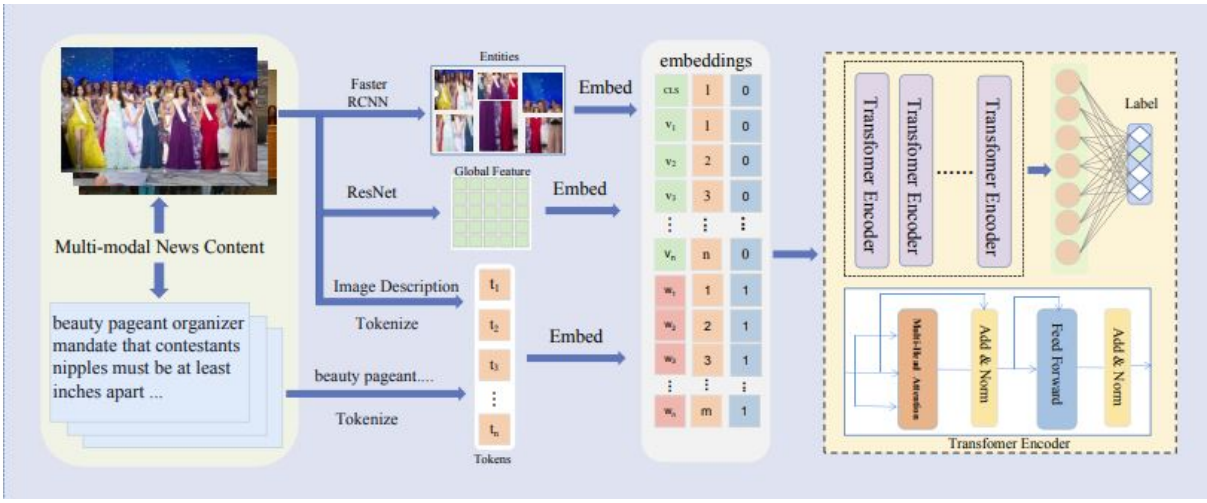

Figure 1: Multimodal architecture [10]

model and an LSTM network to operate on the VGG16 activations, resulting in a final output after a mean pooling layer. The method presented in the article shows improvements over baseline results and the relevancy of using more than one image for gathering visual features.

3 A CRITICAL ANALYSIS

This chapter aims to create an overall picture of the architecture used for each type of task. We will present the algorithms, models, and datasets on which the experiments were carried out and the corresponding research papers used for inspiration.

3.1 Algorithms/Models

Article [12] proposed an architecture based on a hybrid neural network model with multiple layers for extracting characteristics of text context information to ultimately classify news articles. This hybrid model is comprised of a 2-layer Bidirectional Gated Recurrent Unit model and a 3-layer MLCNN model as shown in Figure 2. Text input is vectorized using a trained CBOW model and then forwarded to both BiGRU and MLCNN layers in parallel. The layers are finally merged and return the classification through a softmax layer. Their approach proved state-of-the-art results in classifying Chinese news articles.

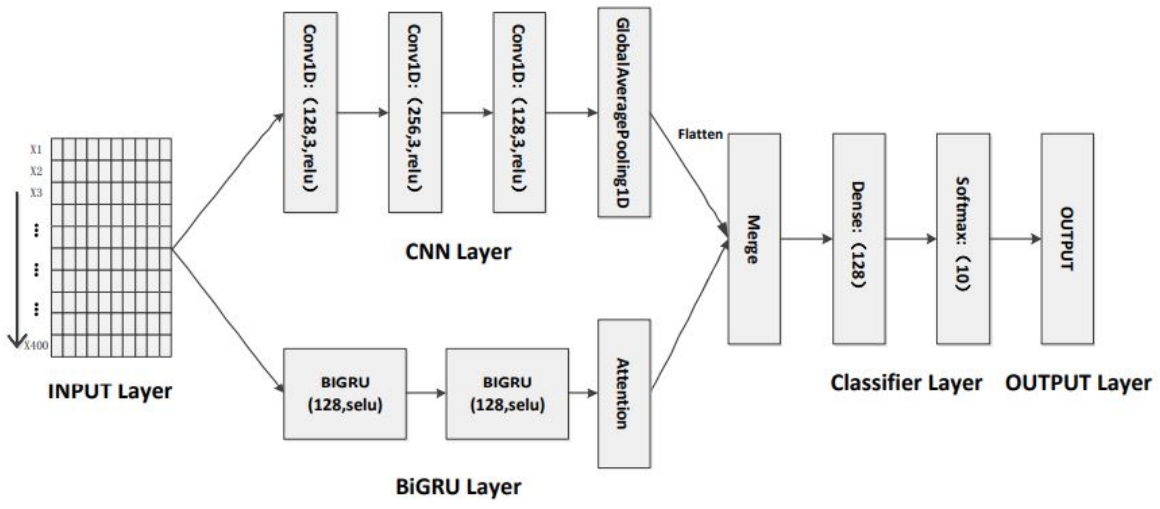

Figure 2: MLCNN & BiGRU ATT model [12]

On the same topic, article [13] also provided useful input in implementing and fine-tuning the BiGRU and CNN architecture used in our experiments on text-only classification, described later in the paper.

Switching to the topic of visual content, articles [2] and [14] were the main sources of inspiration. Transformers in computer vision are a hot topic in today's AI field and the articles mentioned above were a great introduction to this architecture type. They show that CNNs are not

always the answer and yield substantial results. Dosovitskiy et al. [14] provided multiple massively-trained ViT models and variants that can be used with the open-source library HuggingFace [15].

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Figure 3: ViT model variants [14]

The idea of combining textual and visual data came from article [10] presented in the previous chapter. On top of that, article [16] was also a valuable source of information. Zhai et al. presented a method called contrastive-training to align image and text models while still taking advantage of their pre-training. The authors introduced the idea of a locked pre-trained image that teaches the text model to read out good representations for new tasks. In other words, this method allows transferring pre-trained vision models in a zero-shot fashion.

3.2 Datasets

The dataset we focused on with our experiments is Fakeddit [8]. As mentioned before, it is comprised of over 1 million samples of social media posts taken from Reddit. The annotation was done automatically, judging by the subreddit containing the post and the number of positive and negative votes. The labels were also manually verified to ensure the reliability of the data.

There are a lot of other relevant datasets that we will be using in future experiments as well. Some of those are:

- Twitter15 and Twitter16 [17] dataset contains news articles from Twitter, along with their labels and metadata such as users, followers, and followers. This is the only other dataset used in some experiments before testing on the bigger data. This dataset will prove efficient in future mitigation experiments.
- GossipCop [18] dataset consists of over 22000 news articles and the tweet IDs of the users that retweeted the article.
- Kaggle Fake News dataset consists of over 20000 articles labeled as reliable and unreliable.
- Fake News Corpus dataset contains over 9,4 million articles annotated with 11 different labels.

4 PROPOSED SOLUTION

In this chapter, we discuss and summarize the objectives mentioned before and how our experiments will be shaped around them.

The task is to classify a social media post containing some text data and one image into *fake news* (labeled as 1) or *real news* (labeled as 0).

We will present three methods that we explored for each of the following objectives: (1) *Text-only* fake news detection, (2) *Image-only* fake news detection, and (3) *Multimodal* fake news detection.

The dataset used for all the experiments is Fakeddit, working with only multimodal samples and selecting about 7000 total rows without missing data.

1. Text-only fake news detection

For detecting fake news, using only text as our input, we construct a model that resembles the one in [12], modified for our data (Figure 4).
Firstly, we preprocess the text with two steps: (1) cleaning - remove stop words, punctuation marks, and words with a length of 1 and (2) minimizing the vocabulary - lemmatization, all while preserving the original meaning and context.
Secondly, we extract a vectorized word embedding of the cleaned text with a DistilBERT transformer from the HuggingFace library.
Finally, our model will take the embedding as input and forward it to the parallel BiGRU and MLCNN networks. The output of that is then merged with a fully connected layer and a softmax layer.

2. Image-only fake news detection

Switching to image classification, we experimented with fine-tuning a ViT model for classification. We used a pre-trained model, documented in this article [14] (Figure 5), and continued training from the checkpoint.
Before feeding the image into the transformer, it needs to be preprocessed as well. We resize, rescale, and normalize the image. The image is then transformed into a tensor and fed to the model.

3. Multimodal fake news detection

With the two separate classification models, we try to integrate them to get more insight. The method used in the experiments is a rather simple one. We take the embeddings resulting

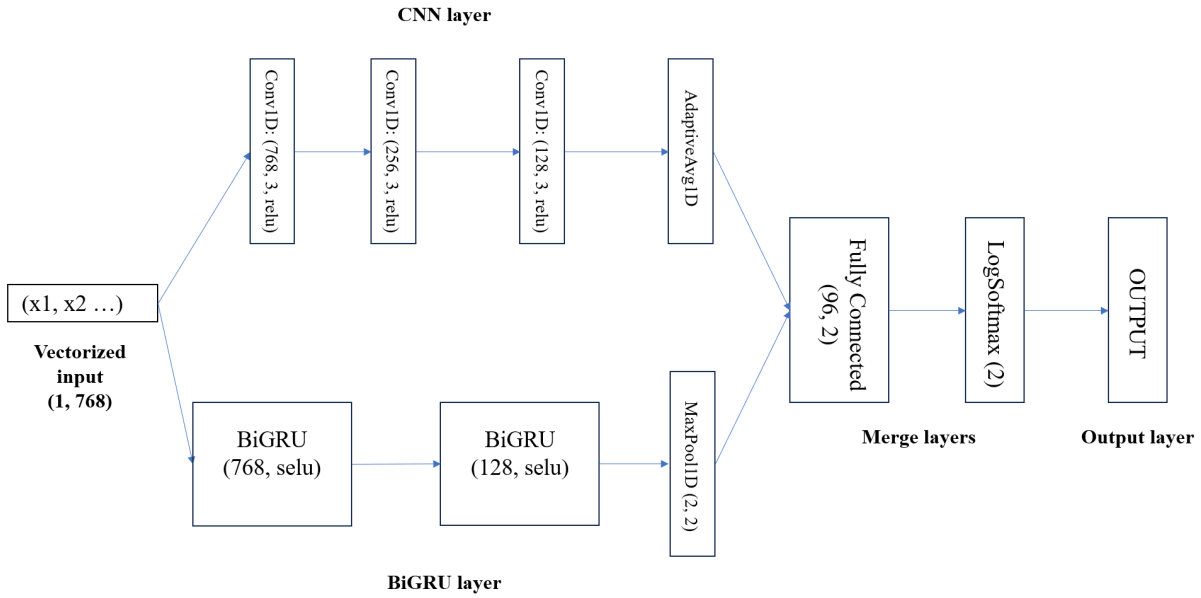

Figure 4: CNN + BiGRU model overview

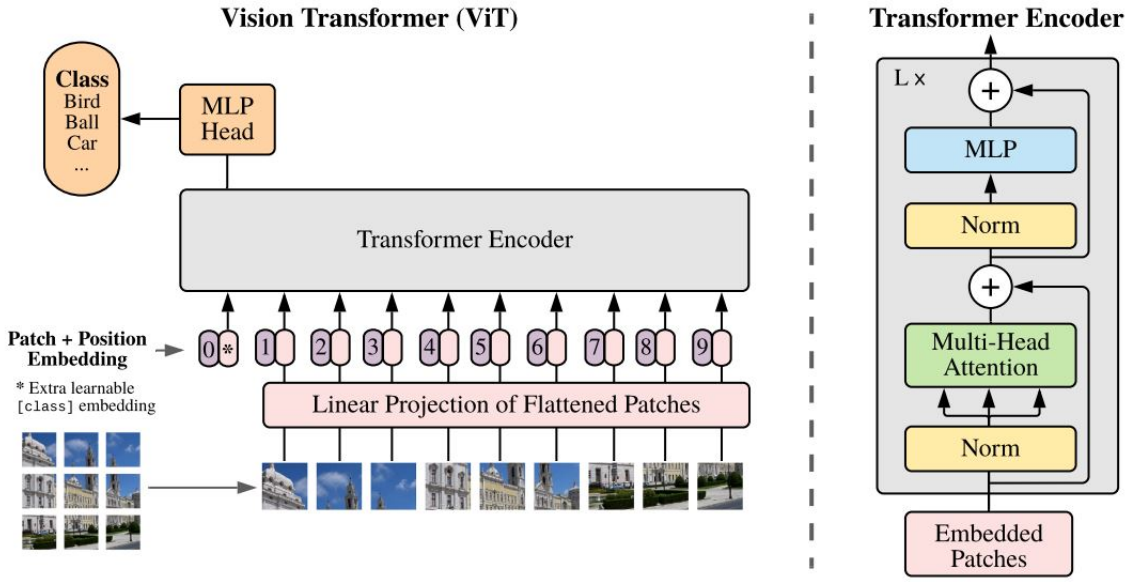

Figure 5: Transformer model overview [2]

from each of the models and we combine them into a softmax layer. In future experiments, we will try to add more depth to this fusion part. As demonstrated in this article [10], adding image-caption techniques could also help identify the similarities and connections between text and image data.

5 PRELIMINARY RESULTS

In this section, we present preliminary results obtained from the implementation of prediction models. These initial findings serve as an early glimpse into the model's performance and set the baselines for further work. Our focus will be on showcasing key metrics and performance indicators, accompanied by a comparative assessment against state-of-the-art models in the field. While these results are provisional and subject to further refinement, they provide valuable insights into the capabilities of our model. The subsequent discussion will delve into specific observations, potential implications, and areas for further optimization, setting the stage for a more in-depth examination of our model's efficacy in comparison to existing benchmarks.

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Twitter15	DistilBERT + BiGRU-CNN	61.23	60.84	61.93	61.38
Twitter16	DistilBERT + BiGRU-CNN	53	52	54	53
Twitter15+16	DistilBERT + BiGRU-CNN	66.36	65.78	65.8	65.79
Fakeddit	ResNet18	60.4	30.2	50	37.7
Fakeddit	BERT + ViT	73.37	73.73	73.15	73.26
Fakeddit	DistilBERT + ViT	74.32	73.77	73.15	73.46

Table 1: Preliminary results

Dataset	Articles	Fake news	Real news	Unverified
Fakeddit	5633	2214	3419	0
Twitter15	1340	332	670	335
Twitter16	740	187	370	181

Table 2: Train datasets detailed

5.1 Datasets details

In all of the experiments, we tested our models on Fakeddit (Figure 6) multimodal samples. We split the data into 6000 rows for training and 600 rows for testing and validating, each. Only rows without missing data were kept, along with the columns for title, image ID, and the 2-way label. To get a better understanding of the dataset, Figure 7 showcases a couple of

examples.

Dataset Statistics	
Total samples	1,063,106
Fake samples	628,501
True samples	527,049
Multimodal samples	682,996
Subreddits	22
Unique users	358,504
Unique domains	24,203
Timespan	3/19/2008 - 10/24/2019
Mean words per submission	8.27
Mean comments per submission	17.94
Vocabulary size	175,566
Training set size	878,218
Validation set size	92,444
Released test set size	92,444
Unreleased set size	92,444

Figure 6: Dataset statistics [8]

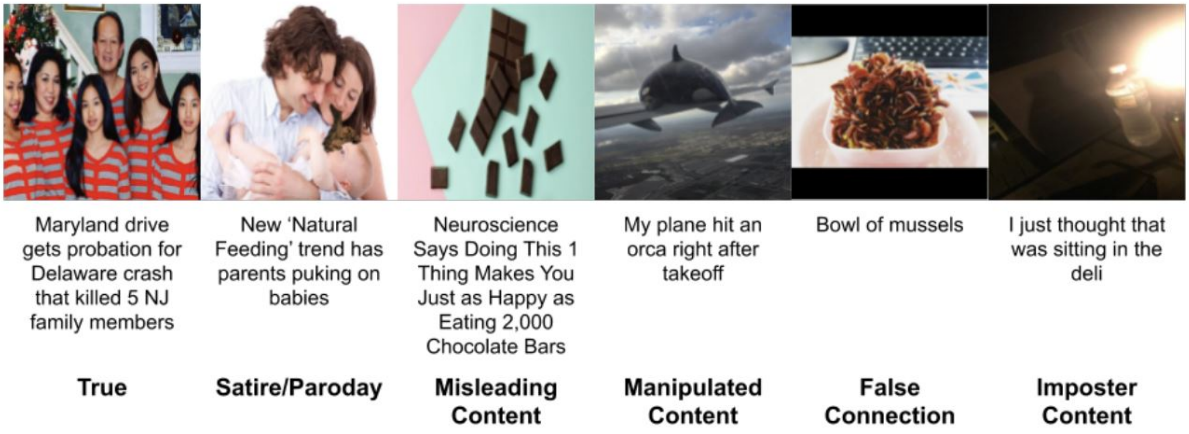

Figure 7: Dataset sample [8]

5.2 Experimental setup

Data loading and stashing were done with pandas data frames [19] and HuggingFace datasets [15]. For text preprocessing, we used NLTK [20] and regular expressions to clean the text, WordNetLemmatizer to extract lemmas, DistilBERT [21] transformer to tokenize and obtain the word embeddings and finally, SciKit-Learn [22] to encode labels.

We used PyTorch [23] to build the networks for the detection module based on text data only. This model was trained for 15 epochs with a batch size of 200, Adam optimizer, and CrossEntropy as the loss function. The learning rate had a value of 0.001 and a weight decay of 0.0005. The input had a shape of (1, 768) after vectorizing with DistilBERT and an output size of (1, 2) with probabilities for each label.
For the image models, we imported the transformer library from HuggingFace and used their library ViTImageProcessor to process the image: resize to 224x224, rescale, and normalize. After that, we fine-tune the ViT for classification over 4 epochs of 16 batches.
In the last step, we take the embeddings resulting from the best models of text classification and image classification and merge them with a softmax layer, resulting in the final predictions.

5.3 Results

Model	Accuracy	Precision	Recall	F1-Score
DistilBERT + BiGRU-CNN	76.73	76.66	75.11	75.88
ViT	74.32	73.77	73.15	73.46
Multimodal	**77.70**	**77.27**	**76.72**	**76.99**

Table 3: Classification results

Results from Table 3 show that the cross-modal method yielded positive results in enhancing the fake news detection task. Moving forward we will continue exploring this side of multimodal detection and further develop the model.

Model	Accuracy	Precision	Recall	F1-Score
MCWDST [4]	76.90	77.27	76.89	76.82
Fakeddit [8]	89.09	-	-	-
DEAP-FAKED [3]	89.55	-	-	-
(BERT+Dense)+Xception [9]	91.87	93.39	93.29	93.25
Multimodal transformers [10]	92.51	93.83	93.74	93.79
Ours	77.70	77.27	76.72	76.99

Table 4: Comparison with state-of-the-art-models

Compared to state-of-the-art models (Table 4), our metrics seem to be quite low, but taking into consideration our physical limitations along with the limitations of the last fusion layer that can be improved, the goal of the experiment was reached - we proved through experiments that the similarity between text and image is relevant to take into consideration for the task of fake news detection in social media.

6 CONCLUSIONS

In this scientific report, we analyzed the importance of combining textual and visual information from social media articles with the purpose of enhancing fake news detection. We experimented with transformer models such as DistilBERT for text embeddings and ViT for visual embeddings, combining two separate models for detecting fake news in text and image with a simple softmax layer. From the results, we observe an increase in accuracy when we combine the modalities, which shows relevance for this area. This topic is still relatively new and not a lot of research has been done towards using multimodal transformers for the detection task. The methods presented in this paper can be improved and more sophisticated approaches are still to be explored.

There is plenty of future work towards a better result in the detection task. The most obvious one is the last layer, where we combine the modalities. Here we could improve by using encoder and decoder techniques for encoding both text and image at the same time and get a dissimilarity matrix, like the ones presented in [16]. In the next research reports, we will also focus on mitigating the spread of harmful content in the social network.

Limitations present themselves when working with larger datasets and processing both text and image. Transformer approaches take up a lot of memory and can be slower than other methods, depending on the computing machine. We also need to take into consideration the social implications of automatically detecting if a user's post is harmful or not. Upholding freedom of speech requires a delicate balance and would need thorough testing before it can be implemented in the real world.

BIBLIOGRAPHY

[1] C.-O. Truică, E.-S. Apostol, T. Ștefu, and P. Karras, "A deep learning architecture for audience interest prediction of news topic on social media," in *International Conference on Extending Database Technology*, 2021.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[3] M. Mayank, S. Sharma, and R. Sharma, "DEAP-FAKED: knowledge graph based approach for fake news detection," *CoRR*, vol. abs/2107.10648, 2021.

[4] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, and P. Karras, "Mcwdst: A minimum-cost weighted directed spanning tree algorithm for real-time fake news mitigation in social media," *IEEE Access*, vol. 11, p. 125861–125873, 2023.

[5] V. Slovikovskaya, "Transfer learning from transformers to fake news challenge stance detection (fnc-1) task," 2019.

[6] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 598–608, 2017.

[7] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," *CoRR*, vol. abs/1908.04472, 2019.

[8] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Marseille, France), pp. 6149–6157, European Language Resources Association, May 2020.

[9] S. K. Uppada, P. Patel, and S. B., "An image and text-based multimodal model for detecting fake news in osn's," *J. Intell. Inf. Syst.*, vol. 61, p. 367–393, nov 2022.

[10] P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, "Multi-modal fake news detection via bridging the gap between modals," *Entropy*, vol. 25, no. 4, 2023.

[11] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal multi-image fake news detection," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 647–654, 2020.

[12] J. Duan, H. Zhao, W. Qin, M. Qiu, and M. Liu, "News text classification based on mlcnn and bigru hybrid neural network," in *2020 3rd International Conference on Smart BlockChain (SmartBlock)*, pp. 1–6, 2020.

[13] Ma, Yuqun, Chen, Hailong, Wang, Qing, and Zheng, Xin, "Text classification model based on cnn and bigru fusion attention mechanism," *ITM Web Conf.*, vol. 47, p. 02040, 2022.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[16] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," 2022.

[17] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, (New York, NY, USA), p. 1867–1870, Association for Computing Machinery, 2015.

[18] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media," 2019.

[19] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference* (Stéfan van der Walt and Jarrod Millman, eds.), pp. 56 – 61, 2010.

[20] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch:

An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.