DistilBERT and RoBERTa Models for Identification of Fake News

A. Kitanovski, M. Toshevska and G. Mirceva
Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje, Skopje, Macedonia
aleksandar.kitanovski@students.finki.ukim.mk, martina.toshevska@finki.ukim.mk, georgina.mirceva@finki.ukim.mk

***Abstract* - The proliferation of fake news has become a significant issue in today's society, affecting the public's perception of current events and causing harm to individuals and organizations. Therefore, the need for automated systems that can identify and flag fake news is critical. This paper presents a study on the effectiveness of DistilBERT and RoBERTa, two state-of-the-art language models, for detecting fake news. In this study, we trained both models on a dataset of labelled news articles and evaluated them on two different datasets, comparing their performance in terms of accuracy, precision, recall and F1-score. The results of our experiments show that both models perform well in detecting fake news, with RoBERTa model achieving slightly better results in overall. Our study highlights the ability of these models to effectively identify fake news and help combat misinformation.**

Keywords - fake news; deep learning; transformer models; DistilBERT; RoBERTa

I. INTRODUCTION

The rise of the internet has brought about a wealth of information at our fingertips, but it has also led to the spread of false information, commonly referred to as fake news. Fake news can have serious consequences, including influencing public opinion, spreading misinformation, and even disrupting elections. As a result, detecting fake news has become an important task in the field of natural language processing (NLP). Recent advancements in NLP and deep learning have led to the development of various techniques for fake news detection. One such approach involves the use of pre-trained language models. Two such models are DistilBERT [1] and RoBERTa [2], these models have achieved astonishing results in various NLP tasks. The current study aims to investigate the performance of these models in detecting fake news. DistilBERT [1] is a distilled version of BERT [3], which is a pre-trained transformer model, while RoBERTa [2] is an improved version of BERT.

The research will be conducted by fine-tuning these models on a fake news detection dataset and comparing their performance on two different fake news datasets. This study is significant in light of the increasing spread of fake news and the need for effective methods to detect and curb its impact. With the growing reliance on social media and the internet for news and information, it has become more important than ever to ensure the accuracy and authenticity of the information we consume. The findings of this study will be of interest to researchers and practitioners in the field of NLP, as well as those concerned with the spread of fake news and its impact on society.

The rest of this paper is organized in the following way. Section 2 presents some related work for solving this task. Section 3 gives detailed description about the approach that is used containing details regarding the datasets that are used for training the prediction models, the preprocessing steps that are made, how the deep learning models, i.e. DistilBERT and RoBERTa are created, as well as the evaluation measures that are used to estimated their predictive performance. The experimental results are presented in Section 4 and are further discussed in Section 5, while Section 6 concludes the paper and gives some directions for further research.

II. RELATED WORK

In recent years, the detection of fake news has been an active area of research. Many studies have used different techniques to address this problem, such as using traditional machine learning algorithms, deep learning algorithms, and natural language processing techniques.

A. Traditional Machine Learning Approaches

Traditional machine learning algorithms, such as Naive Bayes and Support Vector Machines, have been widely used for fake news detection. These algorithms rely on hand-crafted features, such as the frequency of certain words, to classify news articles as true or fake. Ahmed et al. [4] evaluated the performance of six machine learning algorithms including Support Vector Machines (SVM), Linear Support Vector Machines (LSVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD) and Decision Trees (DT). These classification algorithms were applied after extracting TF and TF-IDF feature vectors. The authors in [4] achieved good results with each of the models, with Linear Support Vector Machines achieving the best results.

B. Deep Learning Models

Recently, deep learning algorithms, such as Convolutional Neural Networks and Recurrent Neural Networks, have been used for fake news detection. These algorithms are able to learn complex representations of the articles, and they are better able to handle the context of the words in the articles. However, these methods can be computationally expensive and require large amount of data to train effectively. Saleh et al. [5] compared the performance of several different machine learning

algorithms with three deep learning algorithms, i.e. Recurrent Neural Networks (RNN), Long-Short Term Memory Networks (LSTM) and Convolutional Neural Networks (CNN) optimized for fake news detection. All of their models achieved good results, but CNNs outperformed the rest of the models.

C. Transformers Models

Transformer models have also been used for the task of fake news detection. Mateusz Szczepański et al. [6] have suggested a technique for enhancing the explainability of BERT-based models used in fake news detection. Their approach involves adding an explainability module onto the existing model architecture, thereby avoiding the need for a redesign. A different team of researchers has presented an approach to identify fake news by leveraging the RoBERTa model to detect emotions, which are subsequently employed as features in a Random Forest Classifier to improve the detection accuracy [7]. In [8], another research team has introduced and evaluated an approach utilizing overlapping window strides with multiple transformer models, specifically DeBERTa, RoBERTa, XLM-RoBERTa and BigBird.

This paper focuses on comparing the performance of two specific transformer models, namely DistilBERT and RoBERTa, in detecting fake news. Therefore, our research adds to the existing literature by providing insights into the relative strengths and weaknesses of these models for solving the task at hand.

We want to note that in [9] we already used the RoBERTa [2] model, as well as BERT [3], for solving fake news detection task focused solely on data regarding COVID-19. Namely, in [9], another dataset was used that contains data about tweets related to COVID-19, while in this study we aim to create model for general purpose covering various topics that is a more challenging task.

III. Methodology

A. Datasets

We utilized two datasets in the analysis made in this study. The first dataset was split into three subsets so that 10% of the samples were used for training the models, 20% were used for validation purposes, and the remaining 70% being reserved for testing. We refer to this first dataset as ds-1 and its three subsets are referred to as ds-1 (train), ds-1 (validation), and ds-1 (test) respectively. The reason ds-1 (train) and ds-1 (validation) are only 10% and 20% respectively of the ds-1 dataset is because of the hardware limitations. The ds-1 dataset was sourced from Hadeer Ahmed et al. [10].

The second dataset, referred to as ds-2, was exclusively used for testing purposes. The idea for using this dataset was to being able to test how well the models generalize and if they could be used for new data. The ds-2 dataset was obtained from Kaggle [11].

The class distribution of the classes for all subsets of the ds-1 dataset can be seen in Table 1. It can be seen that the dataset is balanced, which is also a case with its three subsets.

TABLE I. The Distribution of The Classes in the Ds-1 Dataset

Dataset	True	Fake
ds-1 (train)	2044	2321
ds-1 (validation)	4133	4597
ds-1 (test)	14625	15932

TABLE II. The Distribution of The Classes in the Ds-2 Dataset

Dataset	True	Fake
ds-2 (original)	3171	3164
ds-2 (filtered)	3155	3123

The ds-2 dataset has 3164 fake news samples and 3171 true news samples, but some of them were removed due to being very similar to news samples from ds-1, as it will be described in the next subsection. The distribution of the classes in ds-2 before and after this filtering can be seen in Table 2. From Table 2 it can be noted that both the original and the filtered datasets are balanced.

Both DistilBERT and RoBERTa models were fine-tuned on ds-1 (train) using ds-1 (validation) for validation purposes. After fine-tuning the models, they were evaluated on ds-1 (test) and ds-2 (filtered).

B. Preprocessing

To preprocess the data we followed these steps: mapping all letters to lowercase, removing any non-whitespace or non-alphanumeric characters, and removing stopwords from the text. Initially, these preprocessing steps resulted in good results when evaluating on ds-1 (test), but poor results when evaluating on ds-2. Upon further analysis, we discovered that all news articles labelled as "true" in ds-1 began with the same text and that all of them were sourced from Reuters. To address this issue, we added an additional preprocessing step to remove this text after converting all letters to lowercase. This significantly improved the results on ds-2.

To ensure that ds-2 was different enough from ds-1 and to test the models' ability to generalize, we removed all articles from ds-2 (original) that were similar to the articles in ds-1. This was done by computing the tf-idf embeddings of the articles and calculating the pairwise cosine similarity between the samples from the two datasets. Higher value for the cosine similarity indicates dissimilar samples, while low values correspond to high similarity. Through empirical evaluation, we determined that articles with a cosine similarity greater than 0.65 were sufficiently similar and were removed thus obtaining the dataset denoted as ds-2 (filtered). As we can see from Table 2, only a few articles were removed and the class distribution in ds-2 (filtered) remained balanced.

C. DistilBERT and RoBERTa Models

As previously mentioned, in this study we evaluate the performance of two state-of-the-art language models (i.e. DistilBERT [1] and RoBERTa [2]) for solving the fake news detection task.

DistilBERT [1] learns an approximate version of BERT using a knowledge distillation technique [12], [13]. It has only one half of the layers of the original BERT model, thus reducing the number of parameters by 40%. DistilBERT is designed to be smaller and faster than BERT [3], while still retaining much of its accuracy.

RoBERTa [2] is also built upon BERT [3] but has been trained on a larger corpus of text and with a different training procedure. It is trained with dynamic masking, where the masking pattern is generated every time when a sequence is fed to the model, as opposed to static masking in the original BERT implementation where the same training mask was used. RoBERTa has been trained without next sentence prediction objective, with bigger batches over more data and longer sequences.

We have used DistilBERT and RoBERTa implementations from the huggingface library [14]. We fine-tuned both models for 3 epochs using the Adam optimizer. We also used padding and truncation for the tokenizers of both models.

D. Evaluation Measures

We evaluated the models' performance using classification accuracy, precision, recall, and F1-score as evaluation measures. The classification accuracy is an appropriate measure in our case because the datasets are well balanced. These metrics are defined as presented in Eq. (1), Eq. (2), Eq. (3) and Eq. (4).

$$\text{accuracy}=\frac{\text{TP}+\text{TN}}{\text{N}} \tag{1}$$

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{F1-score}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{4}$$

In the preceding definitions, TP denotes the number of true positives (the samples that are correctly classified as positive), TN is the number of true negatives (the samples correctly classified as negative), FP denotes the number of false positives (the samples that are misclassified as positive), FN is the number of false negatives (the samples that are misclassified as negative), while N is the total number of samples.

TABLE III. THE RESULTS FOR DISTILBERT ON DS-1 (TEST)

Evaluation measure	True	Fake
Accuracy	0.99	0.99
Precision	0.99	0.98
Recall	0.98	0.99
F1-score	0.99	0.99

TABLE IV. THE RESULTS FOR DISTILBERT ON DS-2 (FILTERED)

Evaluation measure	True	Fake
Accuracy	0.68	0.68
Precision	0.70	0.67
Recall	0.64	0.72
F1-score	0.67	0.69

TABLE V. THE RESULTS FOR ROBERTA ON DS-1 (TEST)

Evaluation measure	True	Fake
Accuracy	0.99	0.99
Precision	0.99	0.98
Recall	0.98	0.99
F1-score	0.99	0.99

TABLE VI. THE RESULTS FOR ROBERTA ON DS-2 (FILTERED)

Evaluation measure	True	Fake
Accuracy	0.71	0.71
Precision	0.77	0.67
Recall	0.61	0.81
F1-score	0.68	0.74

IV. EXPERIMENTAL RESULTS

In this section, we present the results of our experiments with the DistilBERT and RoBERTa models on the two datasets, ds-1 (test) and ds-2 (filtered). The results for DistilBERT on ds-1 (test) and ds-2 (filtered) are shown in Table 3 and Table 4, respectively. Similarly, the results for RoBERTa on ds-1 (test) and ds-2 (filtered) are shown in Table 5 and Table 6, respectively.

Table 3 and Table 4 show that DistilBERT achieved good performance on ds-1 (test), with high accuracy, precision, recall, and F1-score. However, its performance was not as good on ds-2 (filtered), with lower precision, recall, and F1-score.

Table 5 and Table 6 show that RoBERTa performed similarly to DistilBERT on ds-1 (test), with high accuracy, precision, recall, and F1-score. On ds-2 (filtered),

RoBERTa outperformed DistilBERT, with higher accuracy and F1-score compared to DistilBERT.

Overall, the results suggest that RoBERTa is a better choice for fake news detection compared to DistilBERT, especially considering the fact that the recall is a very important evaluation measure for fake news detection since we want to detect as much of the fake news as possible, namely RoBERTa has a much higher recall for fake news class than DistilBERT on ds-2 (filtered).

We also compared our models with the models obtained in [4], [15], where TF and TF-IDF feature vectors were used in combination with six machine learning algorithms, i.e. Support Vector Machines (SVM), Linear Support Vector Machines (LSVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD) and Decision Trees (DT). We want to note that in [4], [15] the division of the samples into training and test sets was not the same as in our paper. In Table 7 the results from the comparison are provided. From the results it is evident that the DistilBERT and RoBERTa models obtained in this paper are significantly more accurate than the models obtained in [4], [15], which are based on several well-known machine algorithms.

V. DISCUSSION

RoBERTa showed better performance compared to DistilBERT on ds-2 (filtered), especially when considering the recall for the fake news class where RoBERTa was much better than DistilBERT. This is very important since higher recall for the fake news class indicates that the model is more able to detect the fake news. This is especially important in solving fake news detection task in order to reduce the number of unidentified samples that are fake in order to prevent such news to be presented to the end users.

The lower predictive performance obtained on ds-2 (filtered) compared to the case when ds-1 (test) was used could be due to using samples that are focused on other topics for which is harder to determine whether are fake or real based on the data used for training the model. Namely the two datasets could cover texts for very distinct topics.

TABLE VII. THE RESULTS FROM THE COMPARISION WITH SEVERAL EXISTING MODELS

Model	**Accuracy (%)**
Our DistilBERT model	99
Our RoBERTa model	99
SVM	86
LSVM	92
KNN	83
DT	89
SGD	89
LR	89

Our analysis showed that the additional preprocessing step of removing the specific text present at the beginning in the "true" news in ds-1 improved the performance of both models on ds-2 (filtered). These findings suggest that careful preprocessing is crucial for the effective use of these models in detecting fake news. Before applying this preprocessing step, the model identified that specific text that was present in the “true” class and the revealed models were mostly based on that irrelevant knowledge, which was not relevant for ds-2 (filtered) where this text was not present. This way, we became aware that this preprocessing step is required in order to build relevant model, which was later confirmed with the results obtained when adding this preprocessing step. Due to this, we want to note that it is better to assess the model on additional dataset in order to become aware if such patterns are present in the data that should be eliminated in order to build relevant models.

VI. CONCLUSION

In this paper, we focused on solving the task for fake news detection. For that purpose, we utilized the DistilBERT and RoBERTa transformer models in order to create prediction models for solving the task at hand. Two datasets were used in this study. The first one was divided into subsets for training, validation and testing, while the second dataset was solely used for testing purposes.

We compared the performance of the DistilBERT and RoBERTa models for detecting fake news over the two datasets. The results indicate that both models performed well in terms of accuracy, precision, recall, and F1-score. However, RoBERTa outperforms DistilBERT on ds-2 (filtered) regarding the recall for the fake class, thus it better identifies fake news. Additionally, the obtained DistilBERT and RoBERTa models were compared with several models from the literature that are based on several well-known machine learning algorithms, and our models obtained significantly better results.

In conclusion, our study highlights the potential of using transformers such as DistilBERT and RoBERTa in detecting fake news, and underscores the importance of carefully preprocessing the data for improved performance. Further research is needed to explore the use of these models and to evaluate their performance in real-world applications. Besides application of these models in various contexts, also as future work we plan to utilize some other transformer models, which are very popular nowadays since they offer very accurate predictions.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of computer science and engineering at the “Ss. Cyril and Methodius University in Skopje”, Skopje, Macedonia.

REFERENCES

[1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, 2019.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly

optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[3] 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.

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

[5] H. Saleh, A. Alharbi, and S. H. Alsamhi, "Opcnn-fake: Optimized convolutional neural network for fake news detection," IEEE Access, vol. 9, pp. 129471–129489, 2021.

[6] M. Szczepański, M. Pawlicki, R. Kozik, and Michał Choraś, "New explainability method for BERT-based model in fake news detection," Scientific Reports, vol. 11, 23705, 2021.

[7] V. Kolev, G. Weiss, and G. Spanakis, "FOREAL: RoBERTa Model for Fake News Detection based on Emotions," 14th International Conference on Agents and Artificial Intelligence (ICAART 2022), vol. 2, pp. 429-440, 2022.

[8] H. R. LekshmiAmmal and A. K. Madasamy, "NITK-IT NLP at CheckThat! 2022: Window based approach for Fake News Detection using transformers," Conference and Labs of the Evaluation Forum, 2022.

[9] T. Pavlov and G. Mirceva, "COVID-19 Fake News Detection by Using BERT and RoBERTa models," 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO 2022), pp. 312-316, Opatija, Croatia, 2022.

[10] S. S. Hadeer Ahmed and I. Traore, "Kaggle," 2020, accessed: 2022-01-15. [Online]. Available: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset.

[11] A. Motefaker, "Kaggle," 2022, accessed: 2022-01-15. [Online]. Available: https://www.kaggle.com/datasets/amirmotefaker/detecting-fake-news-dataset.

[12] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, "Model compression," 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 535–541, 2006.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[14] Huggingface library, accessed: 2022-01-15. [Online]. Available: https://huggingface.co/.

[15] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," In: I. Traore, I. Woungang, A. Awad (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, ISDDC 2017, Lecture Notes in Computer Science, vol. 10618, Springer, Cham, pp. 127-138, 2017.