2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December, Cox's Bazar, Bangladesh

Detection of Fake News with RoBERTa Based Embedding and Modified Deep Neural Network Architecture

Md. Abdur Rakib Mollah*, Mir Md. Jahangir Kabir†, Monika Kabir‡, Md. Sazid Reza§
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh
University of Technology Sydney, Sydney, Australia†
Murdoch University, Perth, Australia‡
Email: *rakib1703115@gmail.com, †mmjahangir.kabir@gmail.com, ‡monikakabir11@gmail.com, §dihansazid@gmail.com

***Abstract*—The spread of fake news has emerged as a critical challenge in the era of information and digital connectivity. The consequences of misinformation can be profound, affecting public opinion, policy decisions, and even public health. Therefore, detecting and reducing the spread of fake news is an essential issue that requires reliable and precise solutions. Historically, the field of fake news detection has witnessed notable advancements, but several shortcomings have persisted. Many earlier approaches struggled to attain the requisite levels of accuracy. Another formidable obstacle that these earlier approaches encountered was the dynamic and ever-evolving nature of the tactics employed by those who spread fake news. They employ sensational language, which uses phrases with high emotional content to attract readers and elicit a strong emotional response. Recognizing these deficiencies, we present an innovative solution that leverages the state-of-the-art RoBERTa model and a meticulously modified deep neural network architecture. Our approach stands out by not only recognizing the urgency of the fake news detection problem but also by proposing architectural enhancements. It specifically targets the inadequacies of prior methods. We introduce attention mechanisms designed to identify subtle cues indicative of misinformation. The feature extraction techniques capture the nuanced patterns that fake news articles often follow. These architectural refinements make our model extremely effective and achieve an accuracy of 99.76%. The comprehensive evaluations demonstrate that our RoBERTa-based model consistently outperforms previous state-of-the-art fake news detection models, emphasizing the crucial role of advanced language models in combating misinformation.**

***Index Terms*—Natural Language Processing, Transformers, Fake News, RoBERTa, Bidirectional encoder**

I. Introduction

Social media is an integral component of everyday life in the modern world. It provides a conducive environment for establishing contacts, exchanging and producing information, and keeping up with current events. Despite these benefits, it has become more difficult to distinguish between reliable information and low-quality news that is frequently laced with deliberately false material or fake news [1]. Fake news is an unsettling phenomenon that is becoming more and more prevalent. It has the power to affect people's actions and opinions, as well as have a significant impact on both individuals and society as a whole.

During the 2016 United States presidential election, Twitter witnessed a notable surge in the dissemination of fake news, with misleading or entirely false information circulating widely on the platform. These deceptive narratives, often related to candidates and the electoral process, had the potential to sway public opinion, raising concerns about their influence on the election's outcome [2].

Researchers are developing new methods to combat false information on social media due to its widespread dissemination and the potential harm it poses to society. By utilizing Artificial Intelligence (AI) methods, they are attempting to create efficient, automated systems for identifying online bogus news.

In a variety of academic fields, AI approaches have recently shown considerable success. In particular, the use of traditional machine learning techniques or deep learning methodologies for fake news identification has produced encouraging results. When encountering complex real-life scenarios, like text classification and Natural Language Processing (NLP) tasks, traditional machine learning algorithms frequently fall short. Deep Learning (DL) on the other hand has made tremendous strides over time. In addition, whereas DL approaches have the capacity to find significant features inside information automatically, typical ML algorithms require a number of preprocessing steps and feature engineering procedures.

With the introduction of Large Language Models (LLM), significant benefits for the detection of fake news are possible. LLMs excel in content analysis, swiftly identifying inconsistencies, biased language, and unusual patterns that often signal misinformation or disinformation.

Recognizing the gravity of the fake news epidemic and the limitations of previous detection methods, our goal is to introduce an innovative solution that leverages cutting-edge technology to address the shortcomings of existing approaches. In our research, we take advantage of the LLM to extract contextual information from the input texts and classify them with a modified deep neural network. The contributions of our research are as follows:

1. Implementation of a substantial language model and deep

979-8-3503-5901-5/23$31.00 ©2023 IEEE

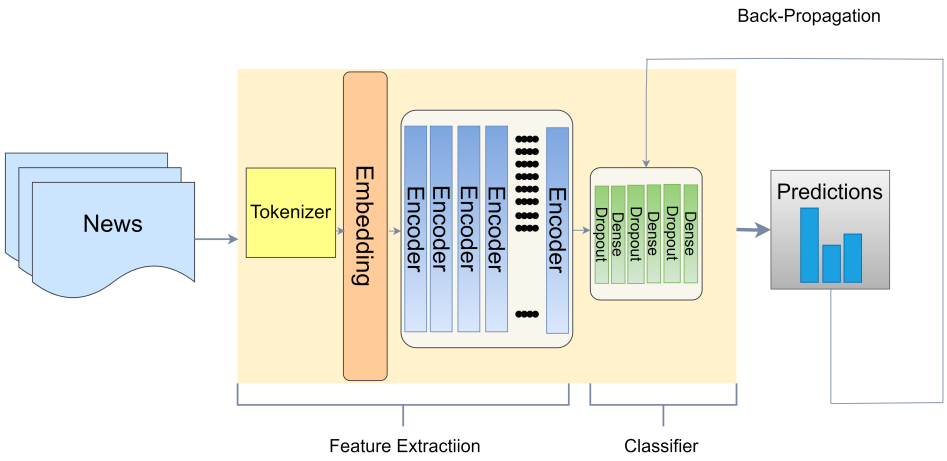

Fig. 1: Proposed Method with RoBERTa Embedding and Deep Neural Network

neural network approach in the study.

2. Focus on identifying fake news through analysis of news content.

3. Utilization of RoBERTa's word embedding and a customized deep neural network structure.

4. Effectiveness of the contributions demonstrated through comparison with previous attempts.

The remainder of this paper is divided into the following sections: we analyze and review previous works in this domain in Section II. Section III presents the design of our suggested system for detecting fake news. In Section IV, the techniques are outlined in detail. It also presents the results and a thorough discussion. Finally, in Section V, we conclude with the findings of the entire research work.

II. Related Works

A paradigm for identifying false information has recently been attempted to be established by many academics. A precision level of 92% was obtained in the work of Ahmed et al. [3]. Term Frequency and Inverse Document Frequency (TF-IDF) features were used by the authors in addition to traditional machine-learning classification methods to identify fake news. In the investigation, six different supervised classifiers were compared and analyzed with the SVM classifier coming out on top.

A proposal appeared with the aim of identifying emerging rumors amid breaking news [4]. Through the use of Word2Vec and the training of a Long Short-Term Memory (LSTM) recurrent neural network, this strategy was based on word embedding. The accuracy that was reached was 79.5%, although the model still has to be improved. A hybrid technique based on an LSTM-CNN model was proposed by the proponents of a different study [5], serving the classification of tweets into either rumors or genuine information. The method's accuracy score of 82% was impressive. The problem with using LSTMs is dominant especially when dealing with relatively small datasets, which may be prone to overfitting.

Another project adopted a hybrid approach for rumor detection that combined Bidirectional Long Short-Term Memory (BILSTM) with several CNN architectures [6]. The model was built using many pre-trained embedded layers, which helped it reach its peak accuracy. A hybrid methodology that included CNN, LSTM, and BILSTM emerged in the field, producing a variety of models that were specifically designed for the identification of fake news based on the interaction between article headlines and article content [7]. The proposed models' highest level of accuracy was 71.2%.

A ground-breaking strategy that used word embedding combined with linguistic features and a voting classification process was revealed [8]. This project's zenith accuracy peaked at a remarkable 96.73%. On the other hand, pursued a strategy that made use of linguistic characteristics, such as stylometric, semantic, and syntactic features. They built a voting system for categorization that was based on conventional machine learning methods [9]. The highest level of accuracy in this situation was 96.36%.

In this research, the limitations of the previous studies are addressed and a different approach is proposed. With this new approach, we successfully gained an accuracy of 99.76% which outperformed all previous efforts.

III. Methodology

A. Dataset Description

We have used the WELFake dataset as it was taken from Kaggle for this investigation [8]. WELFake comprises 72,134 news stories, of which 35,028 are real and 37,106 are fake. Reuters, BuzzFeed Political, Kaggle, and McIntire are four well-known news datasets that the authors combined in order to avoid classifier overfitting and to provide more text data for improved machine learning training. It is crucial to stress that these datasets precisely abide by the standards and norms specified for the creation of unbiased fake news datasets. In essence, these guidelines include the following instructions:

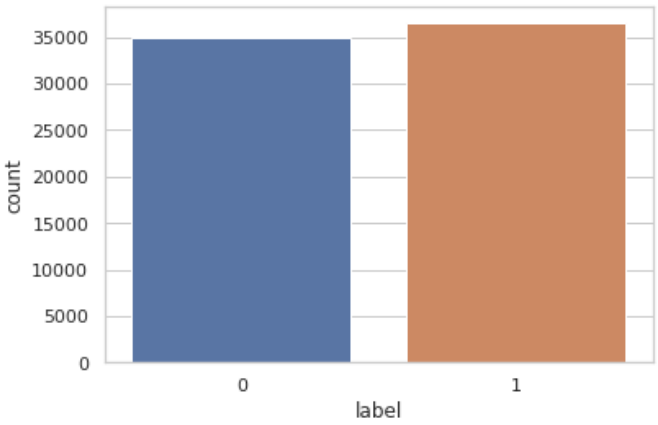

Fig. 2: Distribution of classes

1. Expert Labelling: Each article in the dataset has been painstakingly labeled by subject-matter experts.

2. Diverse Sources for Fake News: A variety of unique sources have been used to compile fake news.

3. Credible Origin of Real News: In contrast, real news pieces have only been obtained from respected and credible journalism organizations.

4. Diverse News Categories: In order to promote diversity, articles have been selected from a wide range of news categories, resulting in the construction of a comprehensive and accurate collection of real news.

The two basic attributes in the WELFake dataset are title and content. Additionally, each article in this dataset has been categorized in the exact same way as follows: 0 indicates false information, while 1 indicates true information.

The dataset has been divided into two distinct sections. 80 percent of the complete dataset, or 57,229 articles, are included in the training dataset. The validation dataset consists of 20 percent of the training data. The test dataset contains 14,308 articles, which makes up the remaining 20%.

B. Pre-trained Language Models

By utilizing their capacity to comprehend intricate contextual linkages in text, pre-trained language models like BERT, RoBERTa, GPT have completely changed the way sentiment analysis is done [10]. Pretraining and fine-tuning are the two basic stages that these models go through. They are exposed to large text datasets during pretraining, which helps them learn the nuances of different languages. By fine-tuning, the model is made specifically for sentiment analysis using labeled data. The model takes a text input and transforms it into a high-dimensional vector representation that captures the subtle meaning. The model is then given a classification layer on top to forecast emotion categories like positive, negative, or neutral. On labeled sentiment data, this classification layer was trained. The process is facilitated by frameworks like Hugging Face's Transformers, which provide straightforward APIs for sentiment analysis using a variety of pre-trained models. However, performance is highly impacted by the model selection and dataset quality.

C. RoBERTa Embedding

The RoBERTa(A Robustly Optimized BERT Pretraining Approach) model for text classification is a development over the earlier BERT language model [11]. The model improves the methods for pretraining by excluding BERT's next-sentence prediction and concentrating solely on the masked language model task [12]. Additionally a pioneer in the use of bigger batch sizes and extended training sequences, it effectively increases its exposure to various linguistic patterns. In terms of classification, there are clear similarities between BERT and RoBERTa. Both involve pretraining a model, adding a task-specific classification layer on top of it and then fine-tuning on labeled data. Similar to previously developed language models, RoBERTa's classification layer adds fully connected neural network layers on top of the underlying model. It is modified to specific classification tasks by these additional layers, which transform its embeddings into predictions. Depending on the difficulty of the job, the architecture may vary; for simplicity, it may consist of a single dense layer with an activation function; for improved performance, it may consist of many dense layers, dropout, and attention mechanisms. A neuron with sigmoid activation for binary tasks or neurons matching classes with softmax for multiclass, producing probability scores, is the layer's ultimate configuration, which is in line with the class count. Task-specific architecture decisions are made in an effort to fully utilize RoBERTa's embeddings for precise prediction.

D. Deep Neural Network

An advanced computational model called a deep neural network architecture is made to recognize complex patterns and features in data [13]. It is made up of numerous interconnected layers of artificial neurons, each with a distinct function. In most cases, the architecture consists of an input layer, hidden layers, and an output layer. Data is entered into the system at the input layer, and features are gradually extracted and transformed at the hidden layers. Artificial nodes, also known as nodes, work together to capture and represent different aspects of the data within these covert layers. The weights and biases that are associated with each connection between nodes are learned during training, allowing the network to adjust and perform at its best.

To reduce overfitting, a common problem in deep learning, dropout layers are frequently used in addition to the core layers. Dropout randomly deactivates a portion of the network's neurons during training, preventing the network from becoming overly specialized and boosting its generalization abilities.

By connecting every neuron in one layer to every neuron in the following layer, dense layers, also referred to as fully connected layers, are essential to the network. As the data moves through the network, these layers enable complex transformations.

TABLE I: Class-specific accuracy, precision, recall, F1-score, and support metrics for the modified architecture

Class	Accuracy	Precision	Recall	F1-Score	Support
0	0.9973	0.9952	0.9990	0.9971	7006
1	0.9979	0.9990	0.9953	0.9972	7302

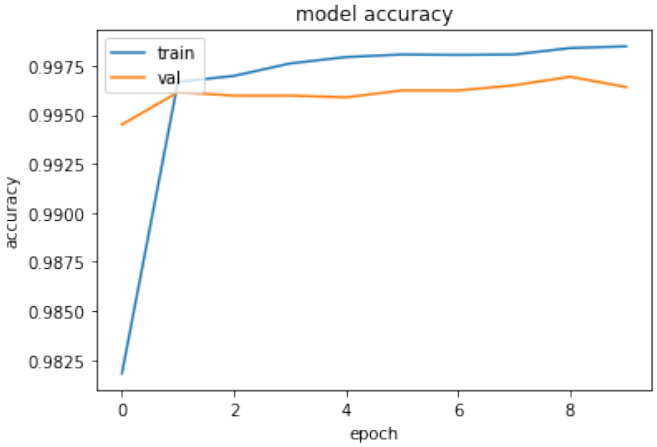

Fig. 3: Training accuracy and validation accuracy of proposed model

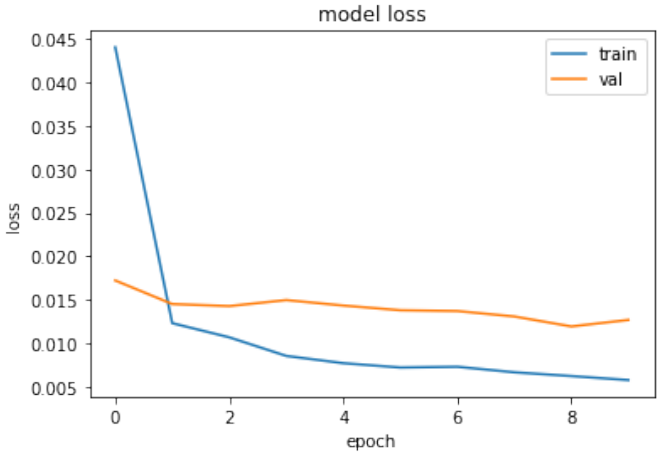

Fig. 4: Training loss and validation loss of proposed model

Deep neural networks are exceptionally good at learning complex data representations from large datasets because of their depth and complexity, which are exemplified by the number of hidden layers and the variety of parameters. This enables deep learning models to excel in a variety of applications, including reinforcement learning, natural language processing, and image and speech recognition.

E. Proposed Architecture

The proposed architecture for the study is a deep neural network composed of several key layers and connections. It starts with two input layers, each designed to accommodate 100-dimensional input data. The two-input architecture, with one input for text data and another for attention masks, enables the neural network to context-aware processing of sequences. The central component of the architecture is the RoBERTa model, which is responsible for processing and extracting features from the input data. This model boasts a substantial parameter count of 124,645,632, showcasing its capability to capture intricate patterns and representations. Following this, a series of dropout layers are introduced to mitigate overfitting, with each dropout layer strategically placed after a dense layer. These dense layers progressively reduce the dimensionality of the data, transitioning from 128 neurons to 64 and eventually to a single output neuron. This architecture demonstrates its potential for complex learning tasks and is poised to excel in various applications, particularly in tasks requiring fine-grained data analysis and prediction. Figure 1 depicts the proposed architecture.

IV. Experimental Analysis

A. Preprocessing

The embeddings from the news texts of the dataset that we obtained would be the classifier's input. But first, stopwords and HTML links were taken out of the texts. The texts were converted using RoBERTa Tokenizer. The tokenizer basically maps words, phrases, or entire sentences to their corresponding dense vector representations after first dividing the texts into words or sequences. Byte-Pair Encoding (BPE) tokenization, a subword tokenization technique, is used by RoBERTa. To make all tokenized sequences the same length, padding tokens are added. It produces an attention mask, which is essentially a binary vector with a length equal to that of the tokenized sequence, after the tokenization and padding processes. The used attention mask separates the relevant input sequence segments that need attention from the irrelevant segments during both the training and inference stages. Token IDs and an attention mask are the final products of the RoBERTa tokenizer, and these items are used as the classifier model's input features for tasks that come after.

B. Experimental Design

In this model architecture, two input layers are employed to accommodate sequences of text data and corresponding attention masks. These inputs are then processed through a pre-trained RoBERTa model, which considers the attention masks while generating contextual representations for each token in the input sequence. A dropout layer with a dropout rate of 0.5 is applied to the contextual representations to prevent overfitting, followed by a dense layer with 128 neurons

TABLE II: Our proposed method in comparison to prior approaches

Work Reference	Overall Accuracy
Merryton et al. [14]	83.00%
Kumar et al. [15]	95.55%
Ribeiro et al. [9]	96.73%
Ouassil et al. [16]	97.74%
Verma et al. [17]	99.01%
Proposed Approach	**99.76%**

and a hyperbolic tangent activation function. Subsequent layers include additional dropout layers (with rates of 0.2 and 0.1, respectively) for regularization and dense layers for feature transformation. The final dense layer, with a single neuron and sigmoid activation, produces binary classification predictions. The model was trained in a batch size of 32, which was the maximum size we could operate due to hardware limitations. 10 epochs were iterated to run the model. For gradient clipping, we used a moderate threshold of 1.0. Momentum and RMSprop optimizations were excluded because the Adam optimizer uses both of their components. Figure 3 and Figure 4 depict the accuracy and loss of the proposed model respectively.

C. Result Analysis

In contrast to previous studies that utilized only a fraction of the dataset, our approach encompassed the entire WELFake dataset, marking a substantial departure from the norm. We adopted an 80:20 dataset split, ensuring the test set's complete independence from the training data. The results were noteworthy: the classifier achieved an accuracy of 99.76% in binary classification. By a significant margin, our approach exceeded the earlier works. TABLE I depicts class-specific accuracy, precision, recall, F1-score, and support metrics for the modified architecture. TABLE II demonstrates the comparison of our work and earlier works.

It's worth noting that the modified architecture, though highly capable, presented computational challenges due to its size, limiting us to a training duration of just 10 epochs. Despite these constraints, our approach demonstrated impressive performance, showcasing the efficacy of our methodology in handling the full dataset and achieving substantial accuracy in binary classification.

V. Conclusion

In this study, we conducted our analysis on the WELFake dataset, a meticulously curated dataset known for its unbiased representation of fake and real news articles. Employing a customized deep neural network (DNN) architecture tailored for binary classification, we achieved an exceptional accuracy rate of 99.76% which is a significant improvement compared to earlier efforts. This remarkable performance underscores the effectiveness of our modified DNN model in accurately discerning between fake and real news articles within the WELFake dataset, setting a new standard in the field of fake news detection. This achievement holds significant implications for improving the reliability and trustworthiness of news content classification systems.

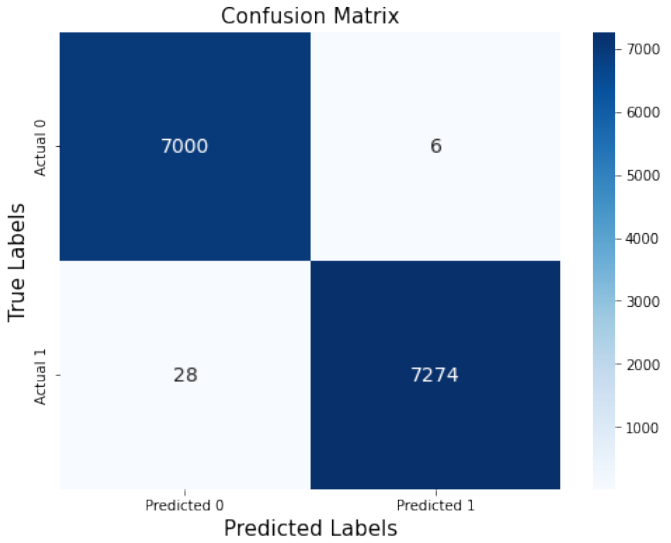

Fig. 5: Confusion Matrix of proposed model for classification

References

[1] J. T. Feezell, "Agenda setting through social media: The importance of incidental news exposure and social filtering in the digital era," *Political Research Quarterly*, vol. 71, no. 2, pp. 482–494, 2018.

[2] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on twitter during the 2016 us presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019.

[3] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*. Springer, 2017, pp. 127–138.

[4] S. A. Alkhodair, S. H. Ding, B. C. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Information Processing & Management*, vol. 57, no. 2, p. 102018, 2020.

[5] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on twitter with hybrid cnn and rnn models," in *Proceedings of the 9th international conference on social media and society*, 2018, pp. 226–230.

[6] M. Z. Asghar, A. Habib, A. Habib, A. Khan, R. Ali, and A. Khattak, "Exploring deep neural networks for rumor detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 4315–4333, 2021.

[7] A. Abedalla, A. Al-Sadi, and M. Abdullah, "A closer look at fake news detection: A deep learning perspective," in *Proceedings of the 3rd International Conference on Advances in Artificial Intelligence*, 2019, pp. 24–28.

[8] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "Welfake: word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, 2021.

[9] J. F. Ribeiro Bezerra, "Content-based fake news classification through modified voting ensemble," *Journal of Information and Telecommunication*, vol. 5, no. 4, pp. 499–513, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] 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.

[12] 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.
[13] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital signal processing*, vol. 73, pp. 1–15, 2018.
[14] A. R. Merryton and M. G. Augasta, "An empirical analysis of fake news detection with nlp and machine learning techniques."
[15] S. Kumar, A. Kumar, A. Mallik, and R. R. Singh, "Optnet-fake: Fake news detection in socio-cyber platforms using grasshopper optimization and deep neural network," *IEEE Transactions on Computational Social Systems*, 2023.
[16] M.-A. OUASSIL, B. CHERRADI, S. HAMIDA, M. ERRAMI, O. EL GANNOUR, and A. RAIHANI, "A fake news detection system based on combination of word embedded techniques and hybrid deep learning model," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, 2022.
[17] P. K. Verma, P. Agrawal, V. Madaan, and R. Prodan, "Mcred: multi-modal message credibility for fake news detection using bert and cnn," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 8, pp. 10 617–10 629, 2023.