Enhancing Fake News Detection with Transformer Models and Summarization

Abdelhalim Saadi
Faculty of Technology, Setif 1 University – Ferhat Abbas, Algeria
abdelhalim.saadi@univ-constantine2.dz (corresponding author)

Hacene Belhadef
Department of Fundamental Computing and its Applications, Faculty of New Technologies of Information and Communication, University of Abdelhamid Mehri – Constantine 2, Algeria
hacene.belhadef@univ-constantine2.dz

Akram Guessas
Department of Fundamental Computing and its Applications, Faculty of New Technologies of Information and Communication, University of Abdelhamid Mehri – Constantine 2, Algeria
akram.guessas@univ-constantine2.dz

Oussama Hafirassou
Department of Fundamental Computing and its Applications, Faculty of New Technologies of Information and Communication, University of Abdelhamid Mehri – Constantine 2, Algeria
oussama.hafirassou@univ-constantine2.dz

Received: 22 February 2025 | Revised: 23 March 2025 and 14 April 2025 | Accepted: 19 April 2025

 DOI: https://doi.org/10.48084/etasr.10678

ABSTRACT

This study evaluates the performance of transformer-based models such as BERT, RoBERTa, and XLNet for fake news detection. Using supervised and unsupervised deep learning techniques, we optimized classification accuracy while reducing computational costs through text summarization. The results show that RoBERTa, fine-tuned with summarized content, achieves 98.39% accuracy, outperforming the other models. Additionally, we assessed AI-generated misinformation using GPT-2, confirming that transformer models effectively distinguish real from synthetic news. We utilized the GPT-2 model instead of more recent models like GPT-4, as our objective was to generate fake news locally and compare it with pretrained models from the same time period.

Keywords-fake news detection; NLP; DL; transformers; RoBERTa; GPT-2; text classification

I. INTRODUCTION

The proliferation of fake news has become a pressing concern in today's digital information age. With the rapid expansion of online platforms and social media, misinformation spreads quickly [1], influencing public opinion, shaping political landscapes, and affecting economic decisions. Traditional methods of fact-checking struggle to keep pace with the volume and speed of misinformation dissemination [2] [3], necessitating automated and scalable solutions. Deep Learning (DL) and Natural Language Processing (NLP) have emerged as powerful tools in the fight against fake news. These technologies enable the development of automated detection systems capable of distinguishing between real and fabricated news articles. Among the most effective models for NLP, tasks are transformer-based architectures, such as BERT, RoBERTa, GPT-2, and T5, which have demonstrated significant improvements in text classification and information retrieval.

The objective of this research is to evaluate the effectiveness of transformer-based DL models in detecting fake news. By leveraging large-scale datasets, we train and fine-tune models to classify news content accurately. Additionally, we explore text summarization techniques to enhance efficiency, enabling models to process large amounts of textual data while reducing computational requirements. Furthermore, we investigate the role of AI-generated content in spreading misinformation by analyzing fake news generated using GPT-2.

II. RELATED WORKS

The increasing spread of fake news has motivated extensive research on automated detection techniques. Early approaches

to fake news classification primarily relied on rule-based systems and linguistic feature analysis. These methods, however, struggled with scalability and adaptability to evolving misinformation patterns [4]. The rise of Machine Learning (ML) and DL techniques has significantly improved fake news detection by enabling models to learn complex linguistic and contextual patterns [5].

Several studies have explored traditional ML techniques such as Support Vector Machines (SVM), Naïve Bayes (NB), and Decision Trees (DT) for fake news classification [6]. While these models achieved moderate success, they often require extensive feature engineering and lack the ability to generalize well across different datasets. DL models, particularly transformers, has provided more robust solutions for handling large-scale textual data with minimal manual feature extraction. One of the most influential transformer models in NLP is BERT (Bidirectional Encoder Representations from Transformers), which introduced bidirectional contextual learning [7], allowing models to understand words based on both preceding and succeeding context. Its fine-tuned versions, such as RoBERTa, have demonstrated superior performance in text classification tasks, including fake news detection. RoBERTa removes the Next Sentence Prediction (NSP) task from BERT's training process and incorporates dynamic masking, leading to more efficient and accurate representations.

Another important approach to fake news detection involves Generative Pre-trained Transformers (GPT-2 and GPT-3). These models, while primarily designed for text generation, have also been investigated for their ability to generate and detect misinformation [8]. Research has shown that AI-generated fake news exhibits linguistic patterns that distinguish it from human-written content, enabling models to classify news authenticity effectively. Additionally, some studies have proposed hybrid models, combining BERT-based feature extraction with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for improved classification [9, 10]. By leveraging contextual embeddings from BERT and sequential dependencies captured by LSTMs, these models achieve higher accuracy in fake news classification tasks.

Recent advancements have explored multi-modal fake news detection [11], integrating textual analysis with image and video processing. Given the rise of misinformation in multimedia formats, future research directions may involve incorporating visual and contextual cues alongside textual analysis for more comprehensive fake news detection systems.

Figure 1 in [12] illustrates the Transformer architecture, which consists of an encoder-decoder structure leveraging self-attention mechanisms and feed-forward neural networks. This architecture forms the basis for advanced NLP models such as BERT, RoBERTa, GPT-2, and T5, which have been extensively used in fake news detection.

III. METHODOLOGY

The methodology of this study is designed to systematically evaluate the performance of transformer-based DL models in fake news detection. This section describes the dataset selection, preprocessing techniques, model architectures, and training procedures employed in the research.

A. Dataset Selection

Two primary datasets were used for training and evaluation:

- Gonzalo/Fake News Dataset: This dataset consists of real and fake news articles, providing a balanced corpus for model training. It includes both headlines and full article texts, allowing models to learn contextual differences between genuine and fabricated news. They contain 40587 articles [13].
- CC_News Dataset: Is a collection of real news articles used to generate synthetic fake news samples using GPT-2. This dataset helps in assessing how well models can differentiate between AI-generated fake news and human-written articles. It contains 708241 English language news articles published between Jan 2017 and December 2019 [14].

B. Data Preprocessing

Preprocessing was a crucial step in optimizing the models for classification accuracy. The following preprocessing techniques were applied:

- Text Cleaning: Removal of special characters, HTML tags, stopwords, and non-alphabetic tokens.
- Tokenization: Splitting text into individual words or subwords using the BERT tokenizer, ensuring compatibility with transformer-based models.
- Summarization: Application of the T5 model to generate concise summaries of news articles, reducing computational overhead while preserving key information.
- Padding and Truncation: Standardization of input sequence lengths to match the transformer model's requirements.
- Label Encoding: Assigning numerical values to categorical labels (1 for real news, 0 for fake news).

C. Model Selection and Architecture

The following state-of-the-art transformer models were evaluated for fake news detection [15]:

- BERT: A bidirectional model pre-trained on vast textual corpora, fine-tuned for classification tasks.
- RoBERTa: An optimized variant of BERT with dynamic masking, enhancing contextual understanding and classification accuracy.
- XLNet: is a transformer-based language model that improves BERT by using a permutation-based training approach. Unlike BERT, which masks words and predicts them independently, XLNet considers all possible word orderings during training, allowing it to better capture context and dependencies. It also retains the benefits of autoregressive models while handling bidirectional context, leading to improved performance on various NLP tasks.

Figure 1 illustrates the complete workflow of the proposed fake news detection system, starting from data acquisition and preprocessing to model training and evaluation. The pipeline

incorporates text summarization using a fine-tuned T5 model, followed by classification with RoBERTa (and BERT and XLNet), and a final comparison between two trained models to determine the best-performing approach.

RoBERTa demonstrated the highest accuracy and efficiency in classifying fake news, making it the primary model for final deployment.

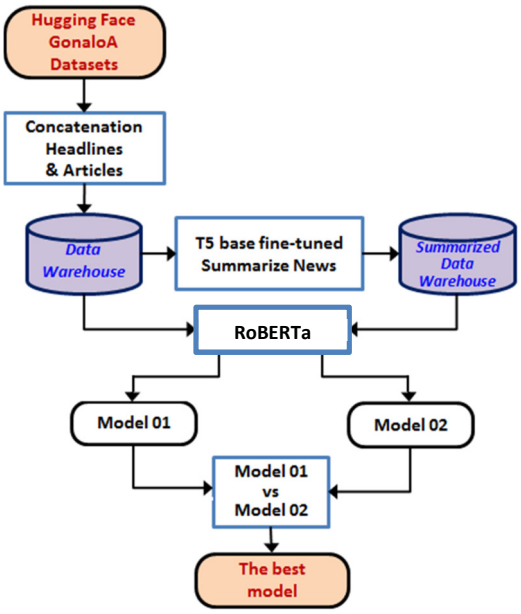

Fig. 1. Fake news detection workflow using transformer models.

D. Model Training and Hyperparameter Tuning

The selected models were fine-tuned using supervised learning on the prepared datasets. Key training configurations included:

- Optimizer: AdamW with a learning rate of 1.25e-6.
- Batch Size: 16, chosen to balance memory efficiency and convergence speed.
- Loss Function: Cross-entropy.
- Early Stopping: Implemented to prevent overfitting, halting training if validation loss does not improve for three consecutive epochs.
- Evaluation Metrics: Accuracy, F1-score, precision, and recall were used to measure model performance.

This methodology ensures a structured and efficient approach to fake news detection, leveraging cutting-edge NLP models and automated text processing. The next section presents the experimental results and performance evaluation of the proposed models.

Figure 2 depicts the various layers of our architecture, starting with the input layer, which represents the contextual word embeddings generated by the RoBERTa model. This is followed by an LSTM layer (768×256) and a classifier consisting of three hidden layers and a binary output layer.

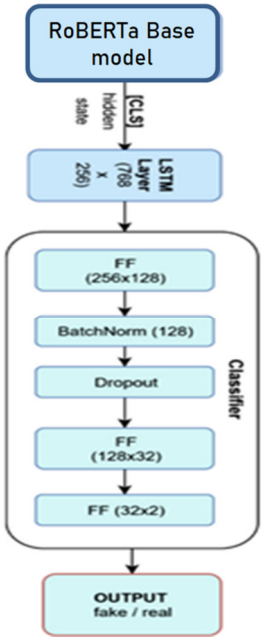

Fig. 2. Architecture of the proposed methodology.

IV. EXPERIMENTS AND RESULTS

This section presents the experimental setup, evaluation metrics, and performance analysis of the transformer-based models employed for fake news detection. The objective is to assess the classification accuracy, computational efficiency, and robustness of various NLP models in identifying misinformation.

A. Experimental Setup

The models were implemented using PyTorch [16] and the Hugging Face Transformers library [17]. Training and evaluation were conducted on a high-performance computing environment, utilizing an NVIDIA RTX 5000 GPU with 32GB RAM. The Paperspace Gradient cloud service was used for model fine-tuning.

The dataset was split into training (60%), validation (20%), and testing (20%) subsets to ensure a fair evaluation of model performance. The training process was conducted over three epochs, with early stopping enabled to prevent overfitting.

B. Evaluation Metrics

The following metrics were used to evaluate the model's performance:

- Accuracy: Measures the overall percentage of correct classifications.
- Precision: Evaluates the proportion of correctly classified fake news instances.
- Recall: Measures the ability to correctly identify fake news cases.
- F1-score: The harmonic mean of precision and recall, providing a balanced assessment of classification performance.

These metrics ensure a comprehensive evaluation of the models' capabilities in detecting fake news.

C. Model Performance Comparison

About the training data, we used two datasets extracted from the GonzaloA/Fake News dataset, Dataset01 that contained the concatenated title and body, and Dataset02 which contained the summarized body (preprocessing).

Model01 is the RoBERTa -base model fine-tuned with Dataset01 and Model02 is the RoBERTa -base model fine-tuned with Dataset02.

TABLE I. RESULTS OBTAINED FROM THE TWO RETURNED MODELS TESTING WITH HEADLINES

Model	Dataset	Pretrained model	Accuracy (%)	F1-score (%)	Training time (min)
Model01	DataSet01	RoBERTa-base	91.58	92	55
Model02	DataSet02	RoBERTa-base	97.87	98.2	14

TABLE II. RESULTS OBTAINED FROM THE TWO RETURNED MODELS TESTING WITH NEWS BODIES

Model	Dataset	Pretrained model	Accuracy (%)	F1-score (%)	Training time (min)
Model01	DataSet01	RoBERTa-base	98.39	98	55
Model02	DataSet02	RoBERTa-base	98.18	98	14

Table I presents the results of testing the models using headlines. Model02 demonstrates a high F1-score of 98%, with an overall accuracy of 97.87%. In contrast, Model01 yields the lowest performance, with an F1-score of 92% and an overall accuracy of 91.58%. Table II displays the results obtained from testing the models using news bodies. Both models achieve an identical F1-score of 98%. However, in terms of overall accuracy, Model01 slightly outperforms Model02, attaining 98.39% compared to Model02's 98.18%.

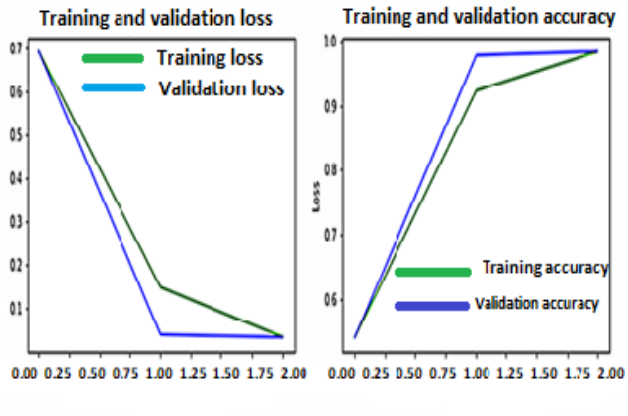

Fig. 3. Training and validation loss/accuracy of the proposed model.

Figure 3 illustrates the training and validation loss, as well as the accuracy evolution of our best-performing model during the training process. The steady decrease in both training and validation loss indicates efficient learning with minimal overfitting. Simultaneously, the accuracy curves show a significant improvement over epochs, with validation accuracy closely following the training accuracy. This confirms the model's strong generalization capability, ensuring high reliability in distinguishing real and fake news.

D. Impact of Summarization on Model Efficiency

To optimize computational resources, T5-based text summarization was applied before classification. The summarized dataset allowed models to process shorter input sequences, reducing training time while maintaining classification accuracy.

- Without summarization, RoBERTa's training time was 55 minutes**.**
- With summarization, the training time decreased to 14 minutes, demonstrating a 75% reduction in computational cost.

This finding highlights the benefit of summarization in improving processing efficiency without compromising accuracy.

E. Analysis of AI-Generated Fake News Detection

To evaluate the ability of transformer models to detect AI-generated fake news, a separate experiment was conducted using BERT and XLNet for unseen data. The results showed that RoBERTa maintained an accuracy of 97.03%, while BERT and XLNet showed respectively an accuracy of 84.21% and 81.66%. The RoBERTa pre-trained model required a longer training time compared to BERT but it achieved better generalization. XLNet, on the other hand, had the highest training time and memory consumption due to its autoregressive-like architecture.

Figure 4 illustrates the precision, recall, and F1-score of the proposed system. The high scores across all metrics indicate strong classification performance, confirming the ability of our model to accurately detect both real and fake news. The high number of correctly classified instances in both classes demonstrates the model's effectiveness. The minimal false positive and false negative values indicate a well-balanced classification capability.

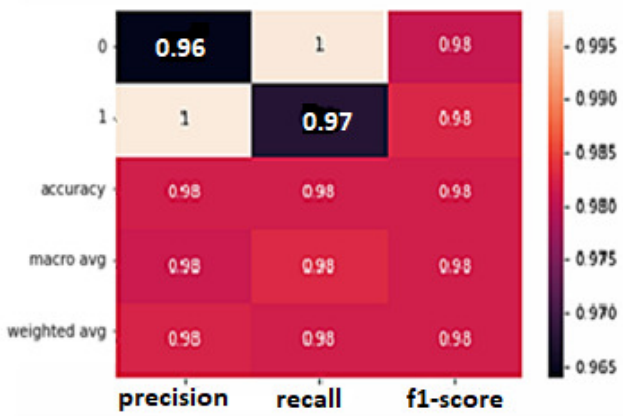

Fig. 4. Classification report of the proposed model.

To evaluate the ability of the proposed transformer model to detect fake news, another experiment was conducted using SVM, NB, and DT for fake news classification for unseen data. The results can be seen Tables III-V.

TABLE III. NB PERFORMANCE

	Precision	Recall	F1-score
0	0.78	0.93	0.85
1	0.91	0.75	0.82
Accuracy			0.84
Macro avg	0.85	0.84	0.84
Weighted avg	0.85	0.84	0.84

TABLE IV. SVM PERFORMANCE

	Precision	Recall	F1-score
0	0.93	0.89	0.90
1	0.89	0.93	0.91
Accuracy			0.91
Macro avg	0.91	0.91	0.91
Weighted avg	0.91	0.91	0.91

TABLE V. DT PERFORMANCE

	Precision	Recall	F1-score
0	0.87	0.61	0. 72
1	0.70	0.91	0.79
Accuracy			0.76
Macro avg	0.79	0.76	0.76
Weighted avg	0.79	0.76	0.76

The results indicate that transformer-based models outperform the SVM, NB, and DT methods, which achieved accuracy values significantly lower than the accuracy achieved by the proposed RoBERTa model.

F. Discussion

The results confirm that transformer-based models significantly enhance fake news detection accuracy. The findings indicate that:

- RoBERTa is the most effective model due to its optimized training and dynamic masking strategy.
- Text summarization reduces training time while maintaining high classification performance.
- AI-generated fake news remains detectable, though continuous improvements are necessary to adapt to evolving misinformation tactics.

These insights highlight the potential of DL models to address the growing challenge of misinformation detection in digital media. The next section discusses the implications of these findings and potential improvements for future work.

V. DISCUSSION

The results obtained in the previous section demonstrate the effectiveness of transformer-based models in fake news detection. This section discusses the key findings, their implications, limitations of the current approach, and potential directions for future research.

A. Key Findings

The experimental results confirm that RoBERTa outperforms other models in fake news classification, achieving the highest accuracy (97.87%) and F1-score (98.2%). The superior performance of RoBERTa can be attributed to its optimized pre-training process, which removes the NSP task and implements dynamic masking, allowing it to generalize better for classification tasks. Additionally, the use of text summarization significantly reduced computational costs without affecting classification accuracy. This indicates that reducing input sequence length through summarization techniques such as T5-based text compression can enhance model efficiency while preserving the ability to distinguish between real and fake news.

The study also demonstrated that AI-generated fake news, particularly content generated by GPT-2, remains detectable using transformer models. RoBERTa maintained 97.03% accuracy when distinguishing between real and AI-generated misinformation. This highlights the potential of deep learning models in counteracting AI-driven disinformation.

We compared our approach with key ML techniques commonly used in fake news detection, such as SVM, NB, and DT. The results demonstrated that our approach outperformed the other methods.

B. Implications for Fake News Detection

The findings of this research have several practical implications for fake news detection systems and digital media regulation:

- Automated Fake News Detection: The deployment of DL-based detection systems in social media platforms and news agencies can enhance the identification of misinformation, reducing its impact on public perception.
- Efficient Computational Approaches: The integration of text summarization can optimize processing time, making real-time detection feasible, especially for large-scale news aggregation platforms.
- Countering AI-Generated Fake News: As AI-generated misinformation becomes increasingly sophisticated, the ability of transformer models to detect these threats provides a first line of defense against AI-driven propaganda.

From a practical perspective, these findings can benefit industries such as e-commerce, customer service, and market research by improving sentiment analysis, automated review classification, and personalized recommendation systems. Additionally, the methodology can be applied to other domains, including healthcare and finance, where opinion mining and text classification are critical for decision-making.

C. Limitations

Despite the promising results, several limitations must be acknowledged:

- Dataset Bias: The models were trained on specific datasets, which may not fully capture the diversity of fake news formats across different languages and regions.

- Generalization Issues: While RoBERTa performed exceptionally well on the selected dataset, its effectiveness on real-world, unseen misinformation needs further validation.
- Lack of Multimodal Analysis: This study focused solely on text-based detection. Fake news often includes multimedia elements (images, videos, deepfake content), which were not considered in this work.

D. Future Research Directions

Our study demonstrates that combining advanced word embeddings (RoBERTa, BERT, and XLNet) with LSTM classifiers significantly improves fake news classification accuracy and robustness compared to traditional ML methods such as SVM, NB, and DT. This finding reinforces the value of DL-based hybrid models for handling uncertain recommendations and complex textual data. To further enhance fake news detection systems, future research should focus on the following aspects:

- Multimodal Fake News Detection: Integrating image, video, and text analysis using multimodal DL techniques to improve detection accuracy.
- Cross-Language Adaptability: Expanding datasets to multilingual news sources to create robust models capable of detecting misinformation in multiple languages.
- Real-Time Deployment: Implementing lightweight DL models for deployment on mobile and web-based applications to facilitate instant fake news verification.
- Our architecture can be used in other fields of natural language processing like sentiment analysis, where the encouraging results in [18] showed the impact of utilizing RoBERTa.

VI. CONCLUSION

The increasing prevalence of fake news poses a significant challenge in the digital era, influencing public opinion, political landscapes, and societal trust. This study explored the effectiveness of transformer-based deep learning models in detecting misinformation, especially RoBERTa which outperformed the other considered transformer models. The integration of text summarization techniques further optimized computational efficiency without compromising detection accuracy.

The results confirm that deep learning models, particularly transformers, provide a powerful approach to fake news detection. The ability of RoBERTa to classify both human-written and AI-generated fake news highlights its potential as a robust misinformation filtering tool. Additionally, the use of GPT-2 for fake news generation and subsequent detection demonstrated that AI-driven misinformation is distinguishable using advanced NLP techniques.

While the findings are promising, several challenges remain, including dataset biases, generalization issues, and the need for multimodal analysis incorporating images, videos, and social media behaviors. Future research should focus on developing cross-lingual fake news detection systems, real-time deployment solutions, and multimodal deep learning approaches to enhance the robustness and applicability of misinformation detection technologies.

In conclusion, this study contributes to the growing field of AI-driven fake news detection by demonstrating the efficiency of RoBERTa and text summarization in classification tasks. The results provide a strong foundation for future advancements in NLP-based misinformation detection, paving the way for more reliable and scalable solutions in combating digital disinformation.

REFERENCES

[1] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017, https://doi.org/10.1257/jep.31.2.211.

[2] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A Stylometric Inquiry into Hyperpartisan and Fake News." arXiv, Feb. 18, 2017, https://doi.org/10.48550/arXiv.1702.05638.

[3] H. F. Villela, F. Corrêa, J. S. de A. N. Ribeiro, A. Rabelo, and D. B. F. Carvalho, "Fake news detection: a systematic literature review of machine learning algorithms and datasets," *Journal on Interactive Systems*, vol. 14, no. 1, pp. 47–58, Mar. 2023, https://doi.org/10.5753/jis.2023.3020.

[4] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, "User Preference-aware Fake News Detection," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, Apr. 2021, pp. 2051–2055, https://doi.org/10.1145/3404835.3462990.

[5] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." arXiv, Sep. 19, 2023, https://doi.org/10.48550/arXiv.1910.10683.

[6] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, "Fake News Classification using transformer based enhanced LSTM and BERT," *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 98–105, Jun. 2022, https://doi.org/10.1016/j.ijcce.2022.03.003.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019, https://doi.org/10.48550/arXiv.1810.04805.

[8] E. Mustafaraj and P. T. Metaxas, "The Fake News Spreading Plague: Was it Preventable?" arXiv, Mar. 20, 2017, https://doi.org/10.48550/arXiv.1703.06988.

[9] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, Apr. 2021, Art. no. 100007, https://doi.org/10.1016/j.jjimei.2020.100007.

[10] N. K. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015, https://doi.org/10.1002/pra2.2015.145052010082.

[11] S. Kumari and M. P. Singh, "A Deep Learning Multimodal Framework for Fake News Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16527–16533, Oct. 2024, https://doi.org/10.48084/etasr.8170.

[12] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Aug. 02, 2023, https://doi.org/10.48550/arXiv.1706.03762.

[13] G0nz4lo-4lvarez-H3rv4s, "G0nz4lo-4lvarez-H3rv4s/FakeNewsDetection." [Online]. Available: https://github.com/G0nz4lo-4lvarez-H3rv4s/FakeNewsDetection.

[14] *S. Nagel, "News Dataset Available," Common Crawl, Oct. 04, 2016. https://commoncrawl.org/blog/news-dataset-available.*

[15] S. Raza, D. Paulen-Patterson, and C. Ding, "Fake News Detection: Comparative Evaluation of BERT-like Models and Large Language

Models with Generative AI-Annotated Data." arXiv, Dec. 20, 2024, https://doi.org/10.48550/arXiv.2412.14276.

[16] J. Jouhar, A. Pratap, N. Tijo, and M. Mony, "Fake News Detection using Python and Machine Learning," *Procedia Computer Science*, vol. 233, pp. 763–771, Jan. 2024, https://doi.org/10.1016/j.procs.2024.03.265.

[17] D. Paper, "Introduction to Deep Learning," in *TensorFlow 2.x in the Colaboratory Cloud: An Introduction to Deep Learning on Google's Cloud Service*, D. Paper, Ed. Berkeley, CA, USA: Apress, 2021, pp. 1–24.

[18] P. Pookduang, R. Klangbunrueang, W. Chansanam, and T. Lunrasri, "Advancing Sentiment Analysis: Evaluating RoBERTa against Traditional and Deep Learning Models," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20167–20174, Feb. 2025, https://doi.org/10.48084/etasr.9703.