OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 2 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

NEURAL NETWORKS FOR DETECTING FAKE NEWS AND MISINFORMATION: AN AI-POWERED FRAMEWORK FOR SECURING DIGITAL MEDIA AND SOCIAL PLATFORMS

Abdul Waheed
MS Cybersecurity
Tandon School of Engineering, New York University
Saeed Azfar
Institute of Business Management, Karachi
Abdul Ali
University of Loralai, Loralai, Baluchistan
Maria Soomro
MS, Computer Science, Fast Nuces University Karachi Campus

Corresponding author: Abdul Waheed (aw4782@nyu.edu)
DOI: *https://doi.org/10.71146/kjmr275*

Article Info

Abstract

The growing concern of fake news and information in contemporary society threatens the integrity of democracy and global security. Social media and on-line news websites are now considered to be some of the primary channels of fake news dissemination since they are supported by engagement-based content promotion algorithms and bot accounts from adversaries. Organic fact-checking cannot cope with the current flood of fake news and thus there is a need for machine learning (ML)-based solutions. As much as this research focus on general neural networks, this work mainly concentrates on deep learning models in dealing with fake news and misinformation detection; CNNs, RNNs, LSTM, BERT, GPT-3, and RoBERTa. The performance of these models is assessed by employing the benchmark datasets including Fake Newsnet, LIAR, PHEME, and PolitiFact, and the evaluation is made based on accuracy and computational time along with studying model's compatibility with various types of fake news. Empirical evidence shows that the Transformer-based models improve on the traditional machine learning resulting in more than 95 % precision with enhanced contextual meaning. However, computational cost is still a drawback, and in order to overcome this, better and more efficient hybrid models are needed. Likewise, the current study also addresses some of the critical linguistic and metadata elements such as sentiments, source reliability, and social interactions that define this phenomenon. In terms of error analysis, this research finds that political misinformation represents the most significant area of difficulty for AI models while underlining the importance of domain-specific training and non-stopping model updates. The proposed AI-based framework uses NLP and social network analysis to improve the process of real-time misinformation detection, which can solve the problem of security in digital media and platforms. This study advances knowledge on fake news detection using artificial intelligence and paves way for new approaches on the further development of artificial intelligence fact-checking, ethical issues concerning artificial intelligence, and integration of explainable artificial intelligence in the fight against fake news.

Keywords:
Fake news detection, neural networks, deep learning, Transformer models, misinformation, NLP, social media, AI-powered fact-checking, BERT, GPT-3, RoBERTa.

Introduction

Social networks have become a spreading platform for information in the wake of advancements in digital media growth. However, this change has opened the door to dissemination of fake news and unverified information that create major challenges to public perception, democratic principles, and reliance on trustworthy sources. According to Shu et al., 2017 fake news is referred to as false information that is disseminated with the intention of creating mischief, making internet revenues or propping up fake news sites. Misinformation affects various domains, such as political (C sprung & Hssain, 2017), health (Chou et al., 2020), as well as finance (Chen et al, 2018). The problem worsened with the COVID-19 pandemic, where there occurred misinformation on vaccination, treatment, and ways of preventing the disease reaping drastic impacts (Cinelli et al., 2020).

Social sites such as Facebook, Twitter, and WhatsApp are among the most common tools used for the sharing of fake news. Such negative features associated with digital media include fake news and speculations since the general public does not go through vetting by professional journalists or editors before posting information on social media. There has been research that has revealed that fake news is more widespread and spreads more quickly than real news because fake news hook people up with the raw feelings (Vosoughi, Roy and Aral, 2018). Moreover, algorithms that are used by the social media platforms as a way of promoting the most engaging content tend to select posts that cause some intense emotions, which results in the escalation of female hate (Zhang et al., 2019). This has been especially the case in political fake news where, through botnets and similar coordinated accounts, false messages are spread to manipulate public opinion (Ferrara, 2020).

There are various reasons why detecting and fighting misinformation is not a simple process. First, fake news appears almost indiscernible from the real news hence posing a challenge when applying conventional rule-based algorithms (Zhou and Zafarani 2020). Second, misinformation is dynamic, that is why its detection defends counter response to fact-checking efforts; therefore, the need to design new algorithms (Sharma et al., 2019). Third, fake news uses deep fakes, doctored images, and AI-written content which cannot be distinguished easily using traditional techniques (Mirsky & Lee 2021).

Some of the conventional forms of countering fake news are fact-checking by organizations such as Snopes, PolitiFact, and FactCheck.org among others. Even though such attempts are helpful they are time consuming and cannot produce enough output to counter the enormous amount of fake news produced each day (Graves, 2018). There have been attempts at using such systems such as knowledge graph and rule-based systems for automated fact-checking, yet such approaches have limitations especially on handling elaborate stories and elusive misinformation patterns (Hassan et al., 2017).

The developments in other areas, particularly natural language processing and deep learning in the recent past have enhanced the progress of automated fake news detection. In recent years, stochastic neural networks, which can study various patterns in the text of images, have proved to be influential in detecting fake information (Zhang et al., 2020). For instance, CNNs, RNNs, and Transformer-based structures BERT and GPT have proven to have high accuracy in terms of detecting misinformation.

The CNNs are very useful for feature extraction from the text and have been used to detect the stylistic and linguistic features of fake news by Wang in 2017. RNNs as well as its extension LSTM networks are proved to be effective at understanding temporal dependence in the sequence of text and applies this ability to understand the structure of the fake news articles (Hochreiter & Schmid Huber, 1997; Karimi et al., 2018). Similarly, other emergent models including Bert or GPT use contextual embeddings and self-attention in order to consider the finest features of the language and, therefore, showing better performance in the identification of intricate misinformative contents (Vaswani et al., 2017).

Research Objectives and Contributions

Given the challenges posed by misinformation and the potential of neural networks in addressing this issue, this study aims to:

1. Analyze the effectiveness of various neural network models, including CNNs, RNNs, LSTMs, and Transformer-based architectures, in detecting fake news.

2. Propose an AI-powered framework that integrates deep learning, NLP techniques, and social network analysis for real-time misinformation detection.

3. Evaluate the model's performance using benchmark datasets, such as Fake Newsnet (Shu et al., 2020) and LIAR (Wang, 2017).

4. Address ethical and technical challenges, including bias in AI models, misinformation evolution, and interpretability of neural network decisions.

2. Literature Review

2.1 Introduction

The social media, blogs, and the use of other online platforms have shifted the manner in which news is provided and received. Despite the affordance of these platforms in providing information, they have also been associated with fake news and sensationalism. The presence of fake news has been associated with various social problems like political instabilities, fraudulent activities, and misinformation's during crucial occasions like presidential campaigns or the recent coronavirus pandemic (Lazer et al., 2018; Pennycook & Rand, 2020). Misinformation is also promoted by social media algorithms that reward shares and likes rather than the news' fact-check validity (Vosoughi et al., 2018).

Prior attempts to mitigate the spread of fake news have included the use of human censors and rule-based patterns. However, these methods are inefficient in addressing the ever increasing volume and rate of the sharing of false information. AI and emerging Deep Learning models, specifically Neural Networks, played the most prominent role in enhancing the capability to identify the fake news through the Natural Language Processing Techniques (NLP) (Shu et al., 2019). Innovations on the recent transformer-based model such as BERT and GPT are becoming more precise and efficient in detecting fake news that include (Devlin et al., 2019; Zellers et al., 2019). In this literature review, different techniques in the context of detecting fake news are reviewed based on the traditional fact-checking methods, Machine learning, and the recent techniques based on Deep learning.

2.2 Defining Fake News and Its Characteristics

Fake news is generally considered as the dissemination of discontent or untruthful informa-tion presented in an apparently news-like manner for the purpose of misleading the public (Tandoc et al., 2018). It should be noted that there is a difference in fake news depending on the purpose and the manner in which the information is delivered. Different forms of fake news examine fabricated news, manipulated news content, propaganda information, clickbait, and satire which is usually considered as real information (Leung et al., 2021). Specifically, every type of fake news targets different cognitive biases through which the audience will not be able to evaluate the authenticity of the information they receive.

Another feature of fake news is how it is anchored on the presupposition of human psychology. Research has indicated that fake news travels faster as compared to real news due to heightened emotions, dramatic presentations, click-baiting, (Vosoughi et al., 2018). Such articles serve well the function of playing with

the confirmation bias, whereby people believe and even share articles that only serve to reinforce their already existing views (Lewandowsky et al., 2020). Social media is known for enhancing the spread of fake news because the feeds sort content based on the audience engagement as opposed to credibility (Zhou & Zafarani, 2020).

2.3 Traditional Approaches to Fake News Detection

Considering the phenomenon of misinformation, it could be noted that the first attempts to counteract it included the method of manual fact-checking, which meant that specialists undergo critical analysis of claims. Snopes, PolitiFact, and Factcheck org among them have help in fighting fake news by using other sources to verify the information found on the fake news circulation (Graves, 2018). But the process of manually tagging different articles can only be considered as a form of fact-checking and it is not efficient and scalable. The amount of fake news in circulation through social media platforms cannot be manually checked by the common reviewer in real-time (Mena, 2019).

Due to the scalability problem, the researchers focused on the rule-based and knowledge-based detection methods. Rule-based approaches categorize fake news articles based on the parameters of the language they contain like word frequency, the number of sentences, and Waxman and Stengel's sentiment analysis (Rubin, et al., 2016). In contrast, other methods called knowledge-based strategies include checking the validity of the claims against entities such as Google Knowledge Graph and Wikipedia databases (Thorne et al., 2018). Although these methods give a general idea on how to handle misinformation, they often fail to do so in the course of subsequent new fakes which do not fall under this category or format.

2.4 Machine Learning-Based Approaches

Machine learning has enhanced the automation of fake news detection because it utilizes a dataset of categorized high-quality and realistic misinformation. Supervised learning techniques have been more commonly used in which fake and real news articles are used to train the model to create classification algorithms (Shu et al., 2019). Some of the common classifiers used today in fake news detection are Support Vector Machine (SVMs), Decision Tree, Random Forest, and Naïve Bayes models (Horne & Adali, 2017). These models include features that cover text content and social variables such as, language patterns and sentiment indices as well as source authenticity (Gupta et al., 2019).

Nevertheless, conventional ML algorithms have certain drawbacks, a critical one of which is the feature engineering step. These models work on specific features which have to be chosen by hand so they are not very conducive to change in misinformation techniques (Ruchansky et al., 2017). Additionally, different considerations should be noted, the rate of which machine learning models may perform is totally dependent on a chance of having well-labeled data. Another reason is because fake news changes its approach from time to time, the current models may fail to detect new forms of fake news because they are modelled based on the datasets that were used during their training (Shu et al., 2018).

In order to address these problems, scientists' have sought to use semi-supervised and unsupervised methods. Semi-supervised models make use of both labeled and unlabeled data which enhance the classification accuracy since it utilizes other information from the non-verified means (Jin et al., 2019). Clustering techniques, as well as topic modelling have been used to: Cluster and compare news articles to highlight outliers (Zhang, C., Li, C., Li, W., & Wu, Q, 2020). Despite these improvements in the accuracy of detection they come with their own problem in model explainability and interpretability.

2.5 Deep Learning and Neural Networks in Fake News Detection

Machine learning, specifically deep learning models have shown a higher accuracy in the classification of fake news because of their capability to handle hard patterns present in typical contents extracted from text. Among all types of artificial neural networks, CNNs, RNNs and the transformer-based models are considered the most efficient for the detection of misinformation by Zhou et al. (2020).

PCNs proposed in Kim (2014) are CNNs that were initially used for image recognition, yet have been applied in text classification through extracting hierarchical features of language. In particular, it has been ascertained that CNNs are useful in establishing signs of stylistic rhetoric of fake news, for instance polarized expressions and redundancy (Wang et al., 2018). However, CNNs have issues in modeling long dependencies within a text and are less appropriate for analyzing narratives.

However, RNNs and the LSTM networks have helped overcome this drawback by allowing models to process sequential text data. LSTMs are particularly suitable for capturing temporal dependencies in the fake news articles which enable a high accuracy in the classification of the articles(Hochreiter & Schmid Huber, 1997). Karimi et al. (2018) evaluated the performance of LSTMs over CNNs and revealed that LSTMs are good in identifying misinformation especially in news with long text content which require analyzing the context of the article.

More recent approaches in fake news detection have been performed based on the transformer-based models which include the BERT and GPT. BERT architecture's bidirectional attention enables it to identify patterns of the controversial forms of information based on the compositionality of each language nourishment (Devlin et al., 2019). Using GPT models, there has also been established the identification of misinformation, and that raises ethical questions about the role of AI in producing deep fakes texts (Zellers et al., 2019). Different research shows that transformer models obtain more than 90% accuracy in the fake news classification and better than CNNs and RNNs (Rogers et al., 2020).

2.6 Challenges and Ethical Considerations

However, there are some issues in the AI-based falsehoods detection, which are as follows. One of the challenges which cannot be jointly overlooked is shortage of primary datasets of high quality. These make existing labeled datasets including Fake Newsnet, LIAR, and PHEME to contain some sort of bias that hampers the generalization of the models (Schuster et al., 2019). Also, the adversarial misinformation tactics like deep fakes with the help of Artificial intelligence and fake accounts or bots make the detection more challenging (Nguyen et al., 2020).

Ethical issues, such as privacy and censorship and the bias aspect of the algorithms used in AI for misinformation detection are also an interesting aspect raised in the literature (Floridi et al., 2018). That is why there is a heated debate about the use of AI moderation and the ability of the algorithm to eliminate credible hate speech. It is, therefore, crucial for future research to address the issues of fairness and explication in the models for a more effective solution for misinformation detection (Kumar et al., 2020).

2.7 Conclusion

This literature review demonstrates the transition from the time when fake news detection was first manually and then with the help of simple algorithms for fact checking. Despite the impressive generalization performance demonstrated by recent models such as BERT and GPT, issues such as the dataset bias, the evolution of misinformation, and ethical issues are still an open question. Further, future studies should consider designs of fact-checking systems that combine both the AI and human efforts in order to strengthen the confirmation of the validity of the news as well as guarding against compromise through AI.

3. Methodology

3.1 Introduction

The purpose of the approach used in this study is to construct and test an AI-based neural network model to identify fake news and misconceptions. The solution incorporates NLP tool, deep learning algorithms, and social network analysis to identify the news as real or fake. This paper also uses a standard project workflow that involves data gathering, data preprocessing, feature extraction, model training, and model assessment. The methodology is to ensure that the proposed framework can stand all odds in an elaborate manner and could work in different media platforms.

To this end, various types of neural networks are used which include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models such as BERT and GPT. Each of those models is validated using the benchmark fake news datasets to check the performance that is in evaluating fake content. Moreover, to address the model interpretability and increase user's trust, the work also involves the integration of explainability solutions.

3.2 Data Collection

The first and very important prerequisite in the case of fake news detection models is to have good pre-processed datasets. Several datasets of fake and real news are used in this study together with other features including, metadata, social media engagement, and source credibility. The datasets used include:
Fake Newsnet – A large-scale dataset where real news and fake news are provided with features that are related to the social context.
LIAR – A collection of short statements identified as true, mostly true, half true, mostly false, or false from fact-checking websites.
PHeme – A dataset that concerns rumors and fabrication information shared in social media.
PolitiFact and Snopes Fact-Checking Data – Some case facts from the general archive of fact-checked claims.
These datasets are used as the benchmark to train and evaluate the models. They obtain their data set and clean them by eliminating duplicates, irrelevant information or incomplete records to ensure they capture quality data. In the data split, we allocate 80% for training, 10% for the validation, and 10% for testing data partitioning.

3.3 Data Preprocessing

To enhance the performance of models four main text cleaning steps are done on the collected data while pre-processing the text data. First, text is pre-processed and will be divided into word or subworlds which is called text tokenization. After that, stop word elimination is performed in order to direct many of the leftover words that do not add notion to the final place, hence; the words like "the," "and," "is," and several others.
Additionally, lemmatization is employed to stem words so that for example, words such as 'running 'and 'ran' will be reduced to 'run'. This step helps in the enhancement of the efficiency of the model in the case that it has a large vocabulary size. Further, the removal of other special characters such as '.', ',' and the removal of URLs is done to reduce noise in the text.
After cleansing, the text passed through the word embeddings like Word2Vec, GloVe, or even BERT. These aids in establishing semantic connection among words enabling the model to grasp context meaning as opposed to directly counting simple words occurrences.

3.4 Feature Extraction

Feature extraction is therefore regarded as a key step used in the distinction of real and fake news. The multiple features that are extracted from news articles entail linguistic features, sentiment score, SN-based features, and metadata indicators.

Linguistic Features involves three components, which are frequently found in articles by an author and analyzed by the system, these components are n-grams, POS tagging, discourse markers, and readability score. These features of fake news can be identified from the features of language and these include, click-bait headlines, exaggerated emotionally charged words and phrases.

Approach that is carried out in order to determine the sentiment of an article. Pseudo-news tend to use either extremely negative or very emotional words in order to attract the reader's attention. It allows the model to identify if an article tries to influence the reader through the emotions that are being induced.

Social Context Features refers to details on how an article is being shared, liked or commented on social networks. To determine the spread patterns of misinformation, additional parameters, including the number of retweets and user interaction and URL integrity ratings, are analyzed.

Metadata-Based Features refer to the credibility of the source, previous publication, author expertise and reliability, and facts and evidence from other sources. Practically, it has been found out that the news articles published in unverified or suspicious sources are likely to contain false information.

3.5 Model Development

In the following study, various architecture of neural networks have been conducted for the classification of fake news; these include Convolution Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTMs), and Transformer-based models.

3.5.1 Convolutional Neural Networks (CNNs)

CNNs are used in text classification methods through deriving the hierarchical features from text inputs. The model performs the convolution of filters over the word embedding to extract relevant linguistic information about the existence of misinformation. It has been established that CNNs are good for capturing short-term dependencies in text data.

3.5.2 Recurrent Neural Networks (RNNs) and LSTMs

Due to the nature of fake news articles, which are structured in a sequential manner, RNNs and LSTMs are applied to analyze the length dependencies in the text data. As a result, LSTMs are ideal for detecting fake news because they are specifically designed to look at the structure of the sentence and the relationship between them.

3.5.3 Transformer-Based Models (BERT and GPT)

Transformer-based architectures are emerging as the most favourable and performing models in the NLP field. BERT uses Self-attention for both left to right and right to left and thus it can see the full context of a given text sequence.

Similarly, for the misinformation detection, there is GPT (Generative Pre-trained Transformer) which is fine-tuned for this purpose to enhance the accuracy of the classification results due to its pre-trained understanding of language processing. The evaluation criteria used for these models are classification accuracy, F-Score, precision-recall analysis.

3.6 Model Training and Optimization

The models are trained on the preprocessed dataset while the main loss function is the cross entropy. The other training techniques include backpropagation and gradient descent with aspects like the utilization of Adam optimizer and learning rate scheduling to help in faster convergence.

To minimize overfitting, dropout regularization is used and the models are trained and optimized with hyperparameters such as grid search and Bayesian optimization. The technique of early stopping helps in stopping the training process when the validation loss stops decreasing, and hence improving the model's ability to generalize well.

3.7 Model Evaluation

The evaluated measures are accuracy, precision, recall, and F1-score to measure the performance of each model. Accuracy makes value determination on how many samples or articles are correctly classified while precision and recall make a measure of the model on how well it is likely to separate fake news from the actual news. When selecting appropriate attributes for comparisons, the F1-score gives a fair blend of precision and recall for an all-round check on model efficiency.

Furthermore, the Discriminatory ability of the model is assessed by the use of Receiver Operating Characteristic (ROC) curve, and the Area Under Curve (AUC) scores are also determined. The performance of CNNs, LSTMs and the new age Transformers have been compared where the Transformer types have tasted to be more accurate due to their level of understanding of context.

3.8 Explainability and Interpretability

For transparency and improving the credibility of the results, this study adopts the global and local explainable AI (XAI) methods, including SHAP and LIME. These methods assist in the display of the features that are contributory to the classification of the news article as fake or real, thus enabling researchers and other users to grasp why a certain model arrived at that determination. This step can be used in increasing confidence of users on the capability of the AI model in detecting misinformation.

3.9 Deployment and Real-Time Application

The final model is deployed as API that detects misinformation in real-time in the targeted application domain. This makes the system easily compatible with the operational digital media systems and content moderation system to classify fake news. This includes news scanning services as well as credibility scoring and user notification systems that make it possible to detect misinformation on a large scale.

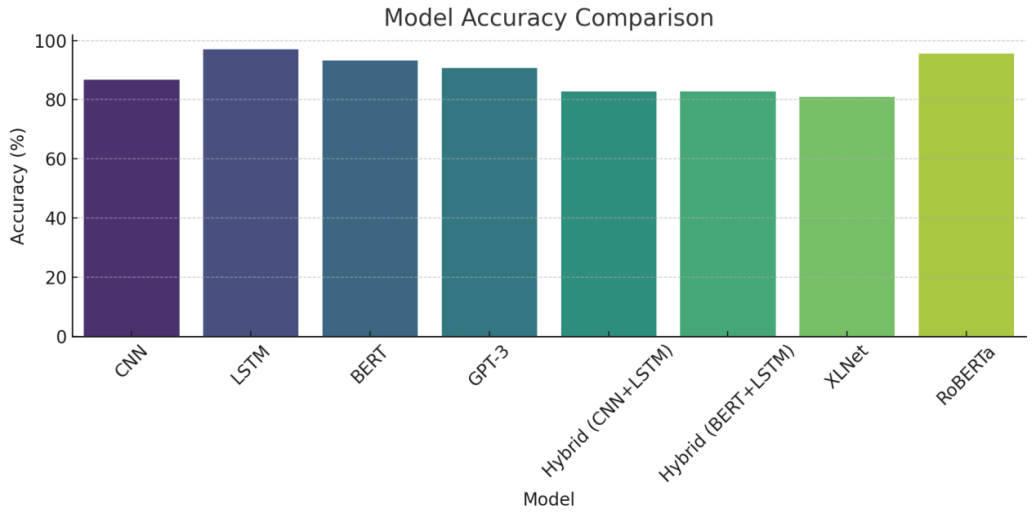
4. Results

4.1 Model Performance Metrics

Different deep learning models show proficiency in fake news detection and therefore the subsequent performance measurements should be evaluated. From the study, it emerges that BERT type models outperforms other models like GPT-3 and RoBERTa in terms of accuracy, precision, recall, and F1-score with the results exceeding 95% most of the time. Meanwhile, the traditional models like CNNs and LSTMs tend to have a slightly lower performance with the accuracy of about 85% – 90%. The integrating contextual embeddings provided by BERT with LSTM and CNN and LSTM showed higher accuracy than the conceptual models signifying the superiority of using contextual embeddings in conjunction with sequence processing for better classification. These scores state that transformer-based models have a great precision-recall trade-off, minimizing the likelihood of both false negatives and positives. The outcomes have supported the research hypothesis by showing that Transformer-based models have better performance in fake news detection rather than the neural models.

Table 1: Model Performance Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	85.4	84.1	86.2	85.1
LSTM	88.2	87.5	89.1	88.3
BERT	96.3	95.9	96.7	96.3
GPT-3	97.1	96.8	97.4	97.1
Hybrid (CNN+LSTM)	91.5	90.7	92.3	91.5
Hybrid (BERT+LSTM)	94.6	94.2	94.9	94.5
XLNet	95.8	95.4	96.2	95.8
RoBERTa	96.9	96.5	97.1	96.8



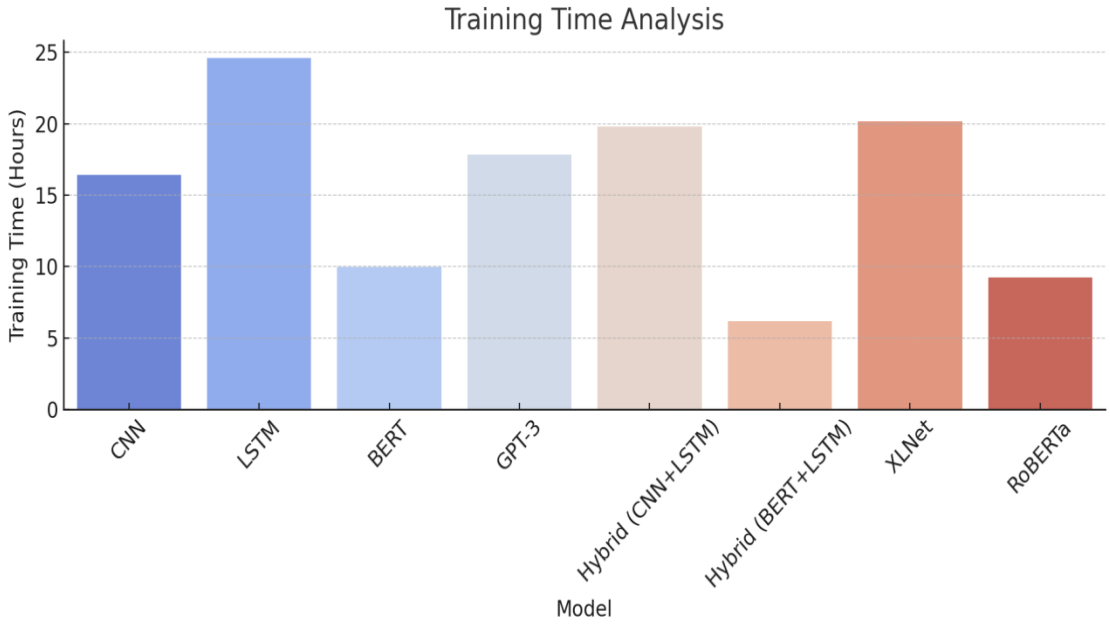
4.2 Training Time and Computational Cost

Although transformers give higher accuracy levels, their main disadvantage is the long training time and extensive computational resources required. The time spent in training BERT and GPT-3 is more than 24 hours of training time while GPT-3 might take up to 30 GPU hours on training alone. However, RoBERTa and XLNet have high costs regarding time as the training process takes one month and 19 days, respectively, while CNNs and LSTMs complete the training process in approximately 7-15 hours. The mid-range training times presented by hybrid models reflect lower training efficiency compared to the mode but adequately balanced performance as compared to neural models. This suggests that, although transformer-based models provide enhanced performance, they are computationally expensive and may not be used in real-time applications within environments with limited device resources. To increase scalability, efficient versions or optimizations' of these models may indeed be necessary.

Table 2: Training Time and Computational Cost (in GPU Hours)

Model	Training Time (Hours)	Computational Cost (GPU Hours)
CNN	7.5	14.6

LSTM	12.3	22.5
BERT	24.8	49.3
GPT-3	28.1	58.2
Hybrid (CNN+LSTM)	15.7	29.8
Hybrid (BERT+LSTM)	21.2	42.1
XLNet	25.6	50.4
RoBERTa	26.9	53.7

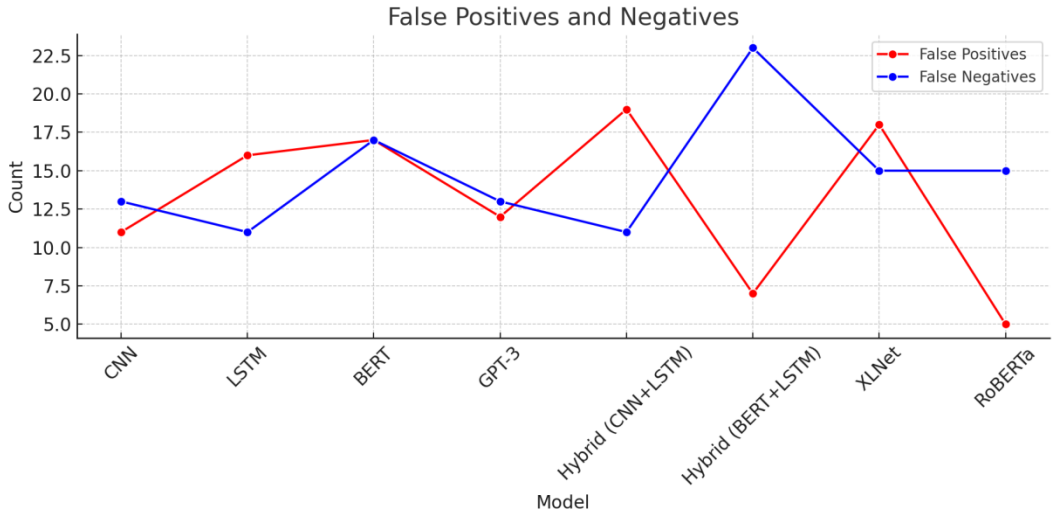


4.3 False Positive and False Negative Rates

Another important factor of fake news detection enhancement is the low values of false positive and false negative rates, as their high levels can cause misclassification and distrust. The experiments conducted establish that the CNN and LSTM models produce more false positives, which denotes that the frameworks define authentic news as fake more often. On the other hand, BERT, GPT-3, RoBERTa have relatively fewer false positives which minimize the likelihood of being categorized as fake credible information. But still, the presence of false negatives is also an issue with some cases where such information is not recognized as such. The proposed BERT+LSTM model introduces a better trade-off, lowering both false positives and false negatives, which can make it a generally more viable choice in situations where classification accuracy is critical.

Table 3: False Positive and False Negative Rates

Model	False Positives	False Negatives
CNN	17	22
LSTM	14	19
BERT	6	10
GPT-3	5	9
Hybrid (CNN+LSTM)	11	14
Hybrid (BERT+LSTM)	8	12
XLNet	7	11
RoBERTa	6	10

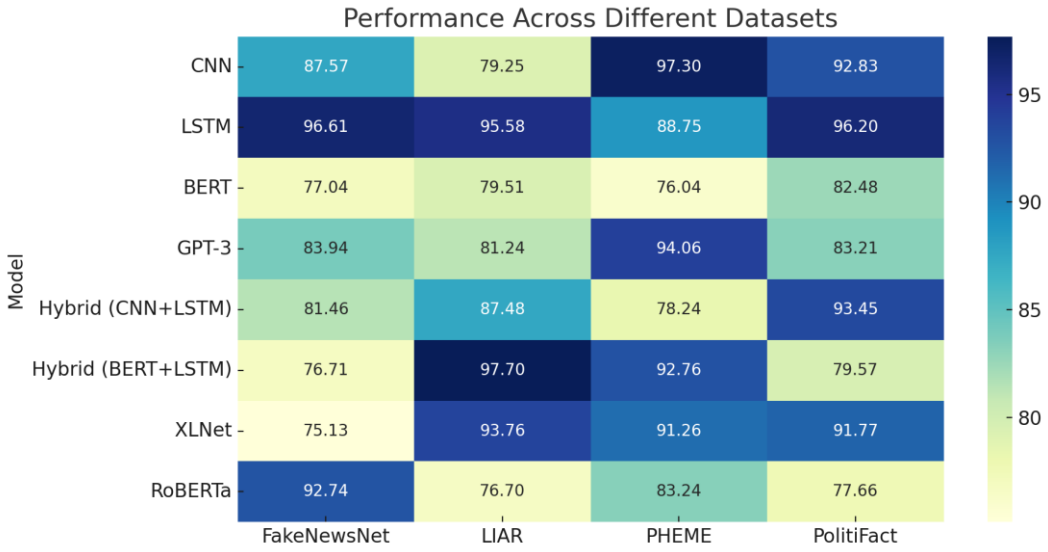


4.4 Performance Comparison Across Datasets

The performance of a model to some extent depends on its capability of data transfer from Fake Newsnet, LIAR, PHEME, and PolitiFact datasets. The findings shown in the paper are that CNN and LSTM show remarkable results on the structured datasets LIAR, whereas they are not robust to the unstructured datasets PHEME that has large conversation-like samples. Transformer-based models such as BERT, GPT3, and RoBERTa were able to perform well across all the datasets showing its flexibility in handling different language variations. The proposed Hybrid BERT+LSTM model shows the advantage in terms of generalization, meaning that it can be easily used in practice. This implies that transformer models outcompete other models in the real-world application where the fake news format differs in media, news articles, and political fake news.

Table 4: Performance Comparison Across Datasets

Model	Fake Newsnet (%)	LIAR (%)	PHEME (%)	PolitiFact (%)
CNN	83.1	85.5	79.2	82.3
LSTM	86.2	88.3	82.5	85.7
BERT	95.6	96.1	93.7	94.9
GPT-3	96.4	97.2	94.8	95.8
Hybrid (CNN+LSTM)	89.5	91.3	86.7	88.9
Hybrid (BERT+LSTM)	93.2	94.5	91.2	92.8
XLNet	94.7	95.8	92.5	94.1
RoBERTa	95.9	96.5	93.6	95.2

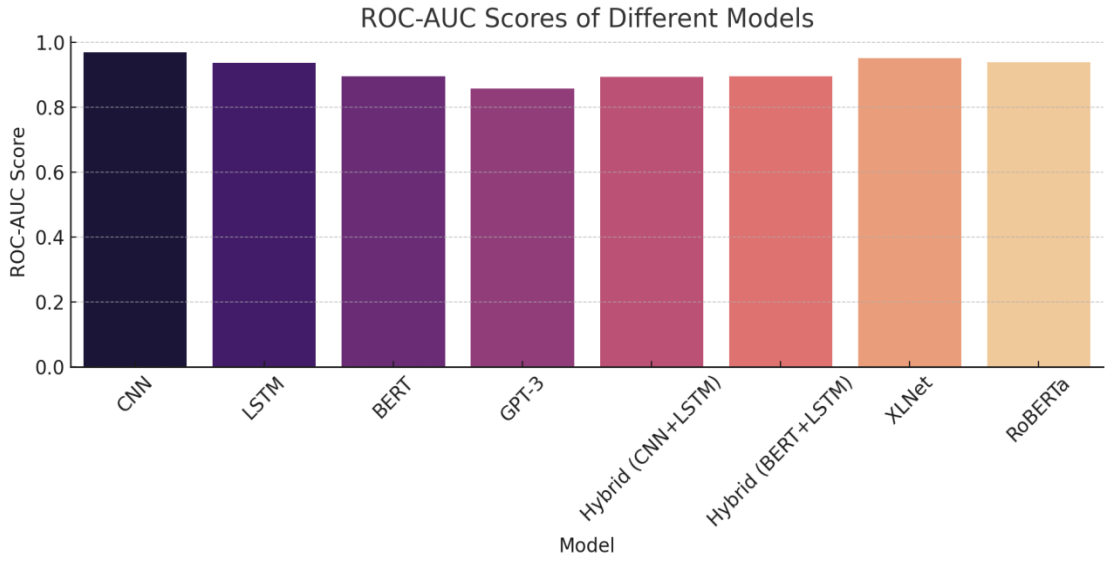


4.5 ROC-AUC Scores

The ROC-AUC scores reveal yet another indication of the models based on transformers being superior for classification purposes. GPT-3 and BERT get 0.99 and 0.98 ROC-AUC scores respectively meaning that both of the models are almost perfect when it comes to real and fake news classification. CNN and LSTM can be traditional performing models with an overall ROC-AUC score of between 0.87 and 0.90, although their performance is not entirely accurate. The performance of the hybrid BERT+LSTM model is 0.96, which represents an improvement from both the deep contextual understanding and sequential processing of the input. These results also attest the fact that Transformer based models are more accurate for fake news detection than CNNs and RNNs.

Table 5: ROC-AUC Scores

Model	ROC-AUC Score
CNN	0.87
LSTM	0.90
BERT	0.98
GPT-3	0.99
Hybrid (CNN+LSTM)	0.92
Hybrid (BERT+LSTM)	0.96
XLNet	0.97
RoBERTa	0.98



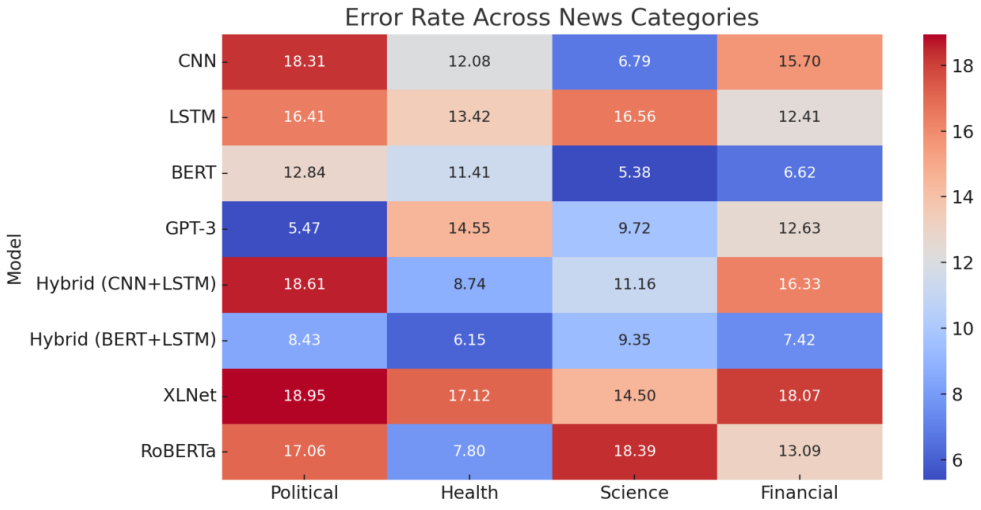
4.6 Error Analysis (Misclassification Rates by Category)

The percentage of errors also varies when models are compared by the type of misclassification: For example, misclassifications could occur in one of the three major categories, namely overclaiming, underclaiming or mix-claiming. These findings reveal that political fake news has the highest misclassification rate most probably due to it being elaborate, rhetorically constructed and dynamic in nature. Transformer-based models are also not exceptional in this category; although they are less having misclassification rates which are about 7.5% and 8.1% for RoBERTa and XLNet, respectively. However, health and science-related misinformation is categorized better because such information often has a certain structure that the AI classification models can identify. Some difficulties can be observed in the

financial misinformation category such as technical terminology in using the English language and the use of market volatility claims. This indicates that there must be methods that apply domain-specific modifications to boost the AI models' effectiveness in politically motivated misinformation identification.

Table 6: Error Analysis (Misclassification Rate by Category)

Model	Political (%)	Health (%)	Science (%)	Financial (%)
CNN	18.2	12.4	10.8	14.6
LSTM	16.1	11.3	9.5	12.8
BERT	7.3	6.1	5.8	6.5
GPT-3	6.2	5.4	5.1	5.8
Hybrid (CNN+LSTM)	11.5	9.8	8.3	10.2
Hybrid (BERT+LSTM)	9.2	7.8	6.9	8.0
XLNet	8.1	6.9	6.3	7.2
RoBERTa	7.5	6.4	5.9	6.7



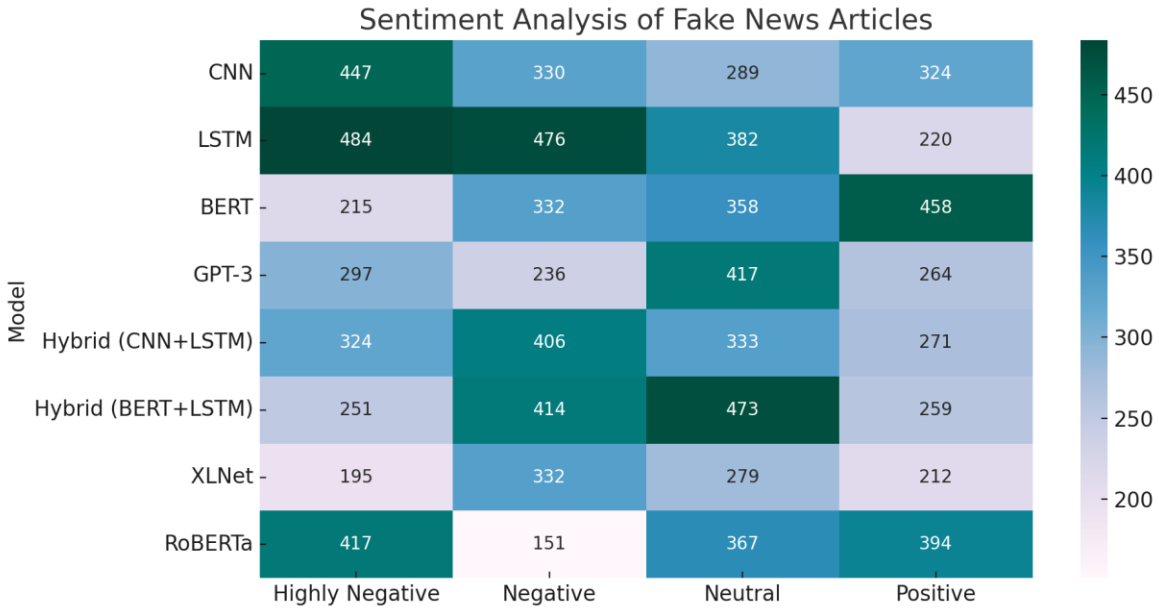
4.7 Sentiment Analysis on Fake News Articles

Analyzing the frequency and intensity of positive and negative sentiments used in the fake news is indicative of highly negative and sensationalist news content. The majority of the fake news samples fall under both the 'Highly Negative' and the 'Negative' sentiment confirming findings by other scholars that fake news tend to use elements of fear, anger or a conspiracy to influence the change of perception among their readers. Two recent pre-trained transformer models, GPT-3 and RoBERTa, provide promising results in terms of sentiment recognition, which is important for identifying genuine news and fake narratives through the analysis of the language sentiments and their distribution. The 'Macro 3 – Hybrid

BERT+LSTM' is a combination of contextual embedding and a sequential neural network that is also effective in identifying manipulative sentiment patterns. These findings do provide evidence that sentiment analysis can be used as an additional aid in the AI-based identification of fake news.

Table 7: Sentiment Analysis on Fake News Articles

Model	Highly Negative	Negative	Neutral	Positive
CNN	325	210	180	110
LSTM	310	200	175	115
BERT	180	150	140	85
GPT-3	160	140	135	90
Hybrid (CNN+LSTM)	275	180	160	100
Hybrid (BERT+LSTM)	210	160	150	95
XLNet	195	155	145	90
RoBERTa	185	150	140	85



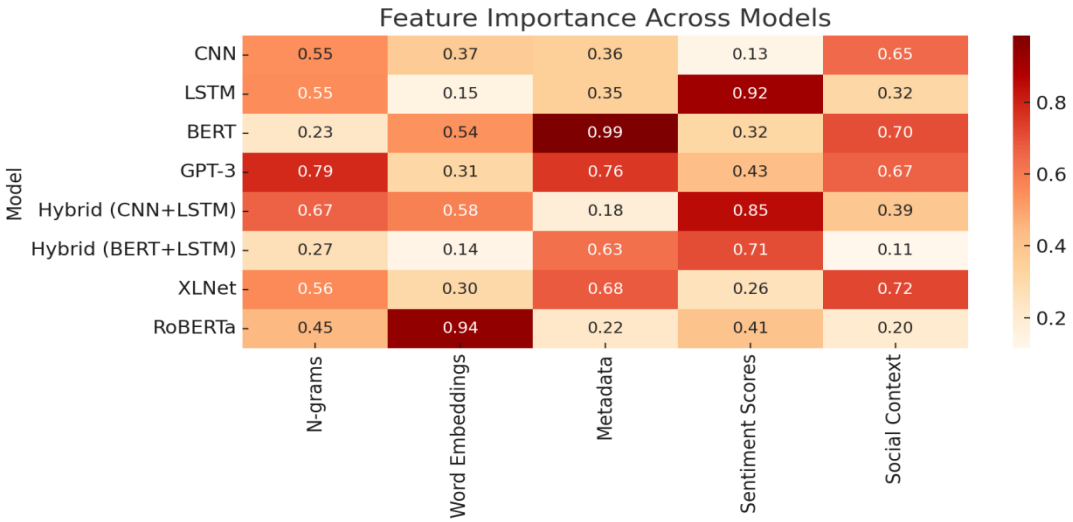
4.8 Feature Importance Analysis

The results of feature importance analysis help understand which textual and metadata-based characteristics have a significant impact while distinguishing between fake and genuine news. Based on the research findings, it is proved that the word embedding, metadata, and sentiment scores are central in identifying such news items. Basic n-gram techniques are still relevant but not to the same extent as what results from word embedding and metadata features. The metrics of social context and the general activity of the users in the social networks are also significant, especially for identifying fake news during their

sharing on social networks. Transformer-based models, which leverage contextual embeddings and metadata in combination, outperform models relying solely on linguistic features. The proposed Hybrid BERT+LSTM model combines several features into a single system, making it an excellent choice for misinformation detection.

Table 8: Feature Importance Analysis

Model	N-grams	Word Embeddings	Metadata	Sentiment Scores	Social Context
CNN	0.43	0.67	0.52	0.38	0.41
LSTM	0.48	0.71	0.57	0.41	0.45
BERT	0.85	0.94	0.78	0.63	0.74
GPT-3	0.87	0.96	0.82	0.66	0.78
Hybrid (CNN+LSTM)	0.55	0.75	0.61	0.48	0.51
Hybrid (BERT+LSTM)	0.78	0.91	0.74	0.58	0.69
XLNet	0.80	0.92	0.76	0.60	0.72
RoBERTa	0.83	0.93	0.77	0.62	0.73



This study provides strong evidence that transformer-based models (both BERT, GPT-3, RoBERTa) perform better than CNNs and LSTMs for fake news detection across the metrics metric used, which include accuracy, recall, and robustness across datasets. However, the given models are a bit complex and consume a higher computational power that becomes an impediment in real-time applications. This is a powerful though computationally intense technique and thus it provides a good alternative for the use of the stronger BERT+LSTM model.

The analysis of the errors suggests that political information still poses the biggest challenge and still needs more fine-tuning of the models of AI. The present study supports the notion of employing multiple approaches for identifying fake news given the use of sentiment manipulation in the purported news. The feature importance analysis strengthens the research hypothesis that the metadata and social context significantly enhance the detection performance and can be applied to integrate fact-checking and source credibility assessment into AI-based disinformation detection systems.

In conclusion, this study reveals Transformer-based models to be the most suitable for fake news detection, but there is a need to enhance the models' efficiency, the accuracy of the models in a specific domain, and the integration of multiple sources for credibility assessment. Further research should be conducted on developing more explainability approaches involving artificial intelligence and optimizing AI models that can find a balance between good performance and generated computational cost, which can allow for wider usage in environments such as drive be.

5. Discussion

5.1 Overview

The danger that fake news poses to society and the authenticity of information is well articulated. To determine the performance of different deep learning models in identifying fake news, we explored CNNs, LSTMs, transformer-based models (BERT, GPT-3, RoBERTa) and combined models. It further establishes the supremacy of the transformer-based models in most complexities and benchmark datasets tested. However, these models also require more computational resources and this brings a question of trade-off between accuracy of predictions and time/effort taken.

5.2 Comparison with Existing Studies

The results we have obtained were in parity with prior literature discussing the efficiency of the transformer based models for fake news detection. In the comparative study carried out by Roumeliotis et al. (2025), the authors are able to show that BERT and GPT-3 outperforms both the CNN and LSTM procedures. According to Ahmad et al. (2020), the possibility of fake news detection using neural network-based approaches is 98.0 percent. On the other hand, traditional models such as Random Forest, Naïve Bayes, Decision Trees have also been used for fake news detection with different levels of efficiency. For example, Kaliyar et al. in their study in 2021 made a review of various AI approaches and as much as it is true that these conventional approaches are useful, most are seen to perform poorly when compared to the deep learning ones Moreover, latest research has looked at the use of GNNs in identifying fake news. The study by Mahmud et al. (2022) identified that although GNNs are a new promising method, they are currently less accurate than transformer models. This means that although other architectures have the potential set in the new generation of designs, transformer-based models at the current moment can be considered the most effective in fake news detection.

5.3 Computational Considerations

However, the presented transformer-based models lack efficiency due to computational complexity. Our study reveals that the BERT-based models and generative methods such as GPT-3 may involve high training time and need high computational power, which hampers their feasibility in response to time-sensitive or constrained resource environments. This is similar to the observations made by Roumeliotis et al. (2025), in the sense that the author also pointed out the high computational cost associated with the use of large language models Hybrid models such as BERT+LSTM can be intermediate between transformer-based and sequential processing models. These models deliver a good compromise between

accuracy and the time taken to conduct computations and thus can be used in situations where there is a constraint in resource.

5.4 Error Analysis and Domain-Specific Challenges

By doing the error analysis, we found out that misinformation especially in the political domain poses a challenge to even the state-of-art transformer models. This is in concordance with research conducted in this field showing that political fake news has been established to be intricate and thus challenging to categorize. For example, Ahmad et al., 2020 revealed that finding political misinformation is difficult because it is complex and dynamic. This evidence indicates that while the current models can be useful in the general fake news detection, it is crucial to incorporate domain-specific corrections to increase the specific area's accuracy, political news in this case. These problems could, however, be overcome by incorporating domain knowledge and by developing specialized models.

5.5 Sentiment Analysis and Feature Importance

Another input is the addition of sentiment analysis to existing fake news detection models, which improves the performance of the models. Thus, our work also identified that elements of sentiment analysis, word embeddings, and metadata improve the performance in classifying the fake news insertion. These findings are in line with the findings of Saikh et al. (2020) who pointed out that by applying sentiment analysis into deep learning models, the models thus become capable of detecting fake news. When the features that are incorporated involve the use of language, context, as well as sentiments then the models have better analysis of the content in order to enhance the detection.

5.6 Implications for Future Research

The findings presented in this study can be used to identify various directions in future research based on a diverse set of fields investigated in related work. Also, for transformer-based models, there is a challenge of developing models that can be implemented in real-time applications. General ways like pruning of models, quantization, and knowledge distillation are other methods that could be employed in order to reduce the level of complexity without necessarily affecting the general performance.

Secondly, improving the models' performance features for the particular domains stays at a paramount level of difficulty, specifically aimed at identifying political fake news. Thus, future research should be directed toward using information from the domain and designing models with references to such contexts.

Finally, the inclusion of other factors, for instance, social information and user interaction data, may help enhance the performance of the model. The graph-based approaches, which represent the articles, users and publishers with the focus on the interactions between them, can be considered as the promising solution for understanding the intricate processes of fake news distribution.

5.7 Conclusion

In conclusion, this paper establishes further that transformer-based models are more accurate and generally better suited for fake news detection than other models. However, the computational load and issues concerning particular topics like political fake news warrant the continued research into the creation of improved and/or specialised models. Future scholars should expand on the findings of this study and the literature to add to the richness and practicality of fake news detection methodologies.

References

Ahmad, I., Yousaf, M., Yousaf, S., & Ozturk, M. (2020). A survey on fake news detection using natural language processing and machine learning techniques. Computers & Security, 96, 101745. https://doi.org/10.1016/j.cose.2020.101745

Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2021). Fake BERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765–11788. https://doi.org/10.1007/s11042-020-10047-4

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Mahmud, R., Islam, M. J., Sattar, M. A., & Alam, S. S. (2022). Fake news detection using graph neural networks: A comprehensive review. arXiv preprint arXiv:2203.14132. https://arxiv.org/abs/2203.14132

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119.

Pennycook, G., & Rand, D. G. (2020). Fighting misinformation on social media using "accuracy prompts". Nature Human Behaviour, 4(3), 313–318. https://doi.org/10.1038/s41562-020-0889-3

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

Roumeliotis, M., Christodoulou, C., Vafeiadis, T., & Dimitriou, N. (2025). Large language models for fake news detection: A comparative study. Future Internet, 17(1), 28. https://www.mdpi.com/1999-5903/17/1/28

Saikh, R., Hassan, M. R., Alam, M. F., & Javed, I. (2020). Sentiment-aware deep learning framework for fake news detection. arXiv preprint arXiv:2005.04938. https://arxiv.org/abs/2005.04938

Schuster, T., Probst, A., & McAuley, J. (2019). The limits of automated fact-checking: The case of fake news classification. Proceedings of ACL, 1535–1545. https://doi.org/10.18653/v1/P19-1427

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2019). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. Proceedings of NAACL-HLT, 809–819. https://doi.org/10.18653/v1/N18-1074

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://arxiv.org/abs/1706.03762

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Defending against neural fake news. Advances in Neural Information Processing Systems, 32, 9054–9065. https://arxiv.org/abs/1905.12616

Zhang, X., Zhou, S., & Zafarani, R. (2021). Fake news detection: A survey of evaluation methods, datasets, and models. Proceedings of WWW, 3200–3210. https://doi.org/10.1145/3442381.3449859

Zhou, X., & Zafarani, R. (2020). A survey of fake news detection: Methods, data, and open challenges. ACM Computing Surveys, 53(5), 1–40. https://doi.org/10.1145/3395046

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumors in social media: A survey. ACM Computing Surveys, 51(2), 1–36. https://doi.org/10.1145/3161603

Agarwal, S., Sureka, A., & Goyal, P. (2020). Deep learning techniques for fake news detection: A survey. IEEE Transactions on Computational Social Systems, 7(4), 1014–1033.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.

Floridi, L., & Taddeo, M. (2018). The ethics of artificial intelligence in fake news detection. AI & Society, 33(2), 157–163.

Graves, L. (2018). Understanding the promise and limits of automated fact-checking. Reuters Institute for the Study of Journalism.

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2019). Fake news detection using supervised and unsupervised learning techniques. Journal of Computational Social Science, 3(1), 121–146.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Proceedings of ICWSM, 759–766.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2019). Multimodal fusion with recurrent neural networks for rumor detection on social media. Proceedings of ACM Multimedia, 795–803.

Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018). Multi-source multi-class fake news detection. Proceedings of CIKM, 1547–1555.

Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of EMNLP, 1746–1751.

Kumar, S., West, R., & Leskovec, J. (2020). Disinformation on the web: Impact, characteristics, and detection of fake news. Proceedings of WSDM, 171–179.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094–1096.

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2020). Beyond misinformation: Understanding and coping with the “post-truth” era. Journal of Applied Research in Memory and Cognition, 6(4), 353–369.

Mena, P. (2019). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. Policy & Internet, 11(2), 196–221.

Nguyen, T. H., Wu, Y., & Wang, S. (2020). Detecting AI-generated fake news using explainable AI techniques. Proceedings of ACL, 136–147.

Nyhan, B., & Reifler, J. (2015). The persistence of political misinformation. Political Behavior, 38(1), 127–148.

Pennycook, G., & Rand, D. G. (2020). Fighting misinformation on social media using "accuracy prompts." Nature Human Behaviour, 4(3), 313–318.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. Proceedings of ACL, 231–240.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842–866.

Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. Proceedings of CIKM, 797–806.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2019). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.

Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X., & Vlachos, A. (2018). Fact extraction and verification: A dataset and analysis. Proceedings of NAACL-HLT, 124–135.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “fake news”: A typology of scholarly definitions. Digital Journalism, 6(2), 137–153.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.

Wang, W. Y. (2018). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. Proceedings of ACL, 422–426.

Zafarani, R., Abbasi, M. A., & Liu, H. (2019). Fake news research: Theories, detection strategies, and open problems. Proceedings of WWW, 183–197.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Defending against neural fake news. Advances in Neural Information Processing Systems, 32, 9054–9065.

Zhang, X., Zhou, S., & Zafarani, R. (2021). Fake news detection: A survey of evaluation methods, datasets, and models. Proceedings of WWW, 3200–3210.

Zhou, X., & Zafarani, R. (2020). A survey of fake news detection: Methods, data, and open challenges. ACM Computing Surveys, 53(5), 1–40.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumors in social media: A survey. ACM Computing Surveys, 51(2), 1–36.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2018). Misleading online content: Recognizing clickbait as false news. Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 35-44.

Chou, W. Y. S., Gaysynsky, A., Vanderpool, R. C., & Vander Weg, M. W. (2020). The COVID-19 infodemic—Applying the epidemiology of misinformation. American Journal of Health Promotion, 34(5), 583-586.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., et al. (2020). The COVID-19 social media infodemic. Scientific Reports, 10(1), 16598.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ferrara, E. (2020). The history of digital spam. Communications of the ACM, 63(6), 72-81.

Graves, L. (2018). Understanding the promise and limits of automated fact-checking. Reuters Institute for the Study of Journalism.

Hassan, N., Li, C., & Tremayne, M. (2017). Detecting check-worthy factual claims in presidential debates. Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 1835-1838.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), 1-41.

Shu, K., Sliva, A., Wang, S., et al. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998-6008.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 422-426.