Twitter Truth: Advanced Multi-Model Embedding for Fake News Detection

Yasmine LAHLOU, Sanaa El FKIHI, Rdouan FAIZI
IRDA Group-ADMIR Laboratory-Rabat IT Center-ENSIAS, Mohammed V University in Rabat, Morocco

***Abstract*—The identification of fake news represents a substantial challenge within the context of the accelerated dissemination of digital information, most notably on social media and online platforms. This study introduces a novel approach, entitled " MT-FND: Multi-Model Embedding Approach to Fake News Detection," which is designed to enhance the detection of fake news. The methodology presented here integrates the strengths of multiple transformer-based models, namely BERT, ELECTRA, and XLNet, with the objective of encoding and extracting contextual information from news articles. In addition to transformer embeddings, a variety of other features are incorporated, including sentiment analysis, tweet length, word count, and graph-based features, to enrich the representation of textual content. The fusion of signals from diverse models and features provides a more comprehensive and nuanced comprehension of news articles, thereby improving the accuracy of discerning misinformation. To evaluate the efficacy of the approach, a benchmark dataset comprising both authentic and fabricated news articles was employed. The proposed framework was tested using three different machine-learning models: Random Forest (RF), Support Vector Machine (SVM), and XGBoost (XGB). The experimental results demonstrate the effectiveness of the multi-model embedding fusion approach in detecting fake news, with XGB achieving the highest performance with an accuracy of 87.28%, a precision of 85.56%, a recall of 89.53%, and an F1-score of 87.50%. These findings signify a notable improvement over traditional machine learning classifiers, underscoring the potential of this fusion approach in advancing methodologies for combating misinformation, promoting information integrity, and enhancing decision-making processes in digital media landscapes.**

Keywords*—*Fake news detection; transformer-based models; text classification; sentiment analysis

I. Introduction

In recent years, the rapid spread of fake news on social media has emerged as a significant challenge to public opinion and even to democratic processes. The widespread dissemination of fake news can cause considerable societal damage, including political polarisation, the erosion of trust in legitimate sources of information and the manipulation of public behavior. In order to meet this challenge effectively, it is essential to implement robust and accurate systems capable of detecting fake news in real time, with the aim of preventing its spread and mitigating its effects.

The traditional methods for the detection of fake news, which often rely on manual verification and heuristic approaches, are no longer sufficient in the context of the current volume and speed with which information is propagated on social media.

In contrast, advanced machine learning techniques offer a promising solution, automating the detection process and improving the accuracy of the results. These techniques facilitate the rapid analysis of large data sets, enabling the identification of patterns and anomalies that may indicate the presence of misinformation. As Lahlou et al. [1] have observed the deployment of machine learning and natural language processing techniques is vital for the identification and categorisation of fake news, given their capacity to handle extensive datasets and extract meaningful attributes. The objective of this study is to utilize advanced methods for optimizing the detection of fake news, particularly on Twitter. Similarly, Shu et al. [2] emphasize the potential of data mining and machine learning in combating the dissemination of fake news on social media platforms.

A number of studies demonstrate the effectiveness of NLP transformation models in identifying fake news, with promising results. In the field of natural language processing (NLP), BERT (Bidirectional Encoder Representations from Transformers) has emerged as a seminal model, distinguished by its capacity to discern intricate contextual nuances in textual data. Developed by Devlin and colleagues [3], BERT has achieved notable levels of accuracy in various Natural Language Processing (NLP) benchmarks through the utilisation of a bidirectional training process and a transformer architectural approach. The utilisation of BERT in the detection of fake news involves the analysis of textual content in an effort to identify any subtle linguistic cues that may indicate misinformation. Another advance in the field of transformer models is represented by the introduction of ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), as proposed by Clark et al. [4]. In comparison to the conventional masked language modelling approach used by BERT, ELECTRA offers a more streamlined training mechanism. This methodology concentrates on the identification of substituted tokens within the textual data, thereby enhancing the model's capacity to discern intricate textual subtleties. In the domain of fake news detection, ELECTRA's discriminative training has demonstrated efficacy in discriminating between authentic and disingenuous content with enhanced precision.

The XLNet approach, as described by Yang et al. [5], integrates autoregressive and auto-encoding methodologies to capture bidirectional contextual dependencies, representing a significant advancement over previous models. By overcoming the limitations of traditional pre-training tasks, the XLNet model

achieves a superior performance in understanding complex linguistic structures and capturing subtle semantic relationships. In the domain of fake news detection, XLNet's capacity to model bidirectional dependencies allows it to identify intricate patterns of misinformation, enhancing the precision and dependability of detection. Recent research has further explored the integration of these transformation models into sophisticated frameworks for the detection of fake news. For instance, Liu et al. [6] illustrated the efficacy of integrating BERT and ELECTRA into a hybrid model. This model combines the respective strengths of both models, thereby achieving enhanced robustness in the detection of various types of fake news.

Despite the advances made to detect fake news, there are still some limitations to existing approaches. Many studies focus only on textual content, neglecting the rich contextual information available from user behavior and social context features as described by Lahlou et al. [7]. Furthermore, a single-model approach may not capture all language features needed for accurate detection [8] [9].

To address these limitations, we propose an advanced, representative approach to detecting fake news by harnessing the collective power of cutting-edge NLP transformative models and additional contextual features. These models, including BERT, ELECTRA, and XLNet, are at the forefront of natural language understanding, each offering unique capabilities to capture complex linguistic nuances [3] [4] [8]. BERT features context-sensitive language modelling using bidirectional transformers, while ELECTRA improves efficiency through a discriminative training approach. XLNet comprehensively captures bidirectional contextual dependencies by incorporating autoregressive and auto-encoding mechanisms.

The objective of this study is to integrate advanced transformer models derived from deep learning with traditional machine learning classifiers for classifying news articles into fake or real categories. A novel approach is introduced that harnesses the capabilities of BERT, ELECTRA, and XLNet transformers to extract contextual embeddings from tweets. These embeddings, in conjunction with sentiment analysis-derived features, serve as comprehensive input features for the classifiers.

The value of our work lies in its innovative approach to integrating state-of-the-art transformer models with traditional machine learning classifiers to tackle the ubiquitous problem of detecting fake news. The study exploits models such as BERT, ELECTRA and XLNet, which are capable of understanding and extracting complex contextual information from news headlines. This in-depth understanding of context is essential for accurately discerning the nuances between true and fake news, which often require a nuanced interpretation of language and sentiment. Furthermore, the incorporation of sentiment analysis-derived features enhances the model's capacity to discern subtle emotional nuances embedded in text, thereby providing a more comprehensive foundation for decision-making in classification tasks.

This study is structured as follows: Section II, titled "Related Works," provides an overview of existing AI techniques and methodologies employed to detect fake news on twitter. Section III delves into the problem formulation, detailing the architectural design and the proposed approach. Section IV outlines the methodology used in this research, describing the experimental setup, data preprocessing techniques, sentiment analysis, and the feature extraction process. It also covers the evaluation metrics used. Section V presents the experimental results, including a detailed comparison of classifier performance on both fused and individual model embeddings, assessing the effectiveness of the fusion approach. Finally, Section VI concludes the research by summarizing key findings, discussing the limitations of the current approach, and proposing potential directions for future work.

II. Related Works

In the past ten years, there has been a great deal of scientific research conducted on the detection of fake news on social media platforms. This study will present an overview of the key scientific research models developed to detect fake news on Twitter. The aim is to provide a comprehensive summary of these models. A threefold classification of the models in question is proposed: traditional machine learning models, deep learning models and transformation models.

With regard to machine learning, varieties of techniques are available for the differentiation between true and fake news on Twitter. These techniques include, but are not limited to, logistic regression, long-term memory, K-mean, support vector machine (SVM), random forest (RF), and Naive Bayes (NB). These techniques employ data pre-processing, feature extraction and sentiment analysis in order to enhance the accuracy of classification models. The linear SVM classification algorithm, which employs TF-IDF feature extraction, demonstrated the highest accuracy of 99.36% [10] [11]. Similarly, Raja [12]achieved a high level of accuracy (93%) using TF-IDF and an SVM classifier. Srinivas [13] employed TF-IDF in conjunction with MNB, RF, SVC and LR classifiers, achieving an accuracy of 79.05%.

Recent advancements in automated fake news detection have increasingly emphasized the integration of contextual features, including temporal patterns, social context, and cross-platform data. The analysis of temporal patterns, particularly in terms of how information spreads over time, can provide invaluable insights into the veracity of the content in question. For example, the rate and timing of tweet propagation frequently display anomalous spikes when fake news is disseminated.

The incorporation of contextual data into the process of feature extraction represents a pivotal stage in aligning the core intent of tweets with their content. This approach facilitates the enhanced detection of misinformation propagation on Twitter [14]. Zhou and Zafarani [15] demonstrated that temporal dynamics, when analysed with machine learning models, enhance the detection of fake news by capturing the evolving nature of misinformation. Furthermore, the utilisation of social context is an efficacious methodology for the identification of misinformation. The analysis of social signals, including user interactions, follower networks and retweet patterns, can provide valuable additional information that can be used to distinguish between genuine and fabricated content. As demonstrated by Shu et al. [16], machine-learning models trained on social network data are capable of effectively assessing the credibility of users disseminating information,

which facilitates improved detection of fake news. The study demonstrated that incorporating social context into feature extraction enables models to not only detect fake news but also trace its propagation paths, thereby providing valuable insights into the dynamics of misinformation spread.

Shetty et al. [17] put a comprehensive framework that incorporates linguistic features, user engagement models, and network analysis, thereby demonstrating effective detection of fake news. Bhogi et al. [18], employed a range of machine learning (ML) techniques, including natural language processing (NLP), classification algorithms, and anomaly detection, achieving an accuracy of 93% with the passive-aggressive classifier.

Another promising approach to the automatic detection of fake news is cross-platform analysis. Misinformation frequently disseminates across a multitude of social media platforms; thus, cross-platform analysis can discern patterns that may evade detection on a single platform. In a related study, Nguyen et al. [19], developed a framework that integrates data from various social media platforms with the objective of detecting inconsistencies in the manner in which news is presented and shared. The results indicate that machine-learning models incorporating cross-platform data are more resilient and effective in identifying fake news, as they are capable of detecting anomalies characteristic of coordinated misinformation campaigns.

Collectively, these studies demonstrate the potential of ML in detecting fake news on social networks.

The transition from machine learning to more sophisticated models, including those based on deep learning, has been instrumental in advancing the field. This has been achieved through the exploitation of the abilities of convolutional neural networks (CNNs), recurrent neural networks (RNNs) and long-term memory networks (LSTMs). Such models can successfully identify and analyse complex patterns within tweet content, user behavior and propagation dynamics. For example, Manaskasemsak and Rungsawang [20].proposed a deep neural network model that leverages tweet content published time and social graph features for the detection of fake news sources on Twitter with high accuracy.

Furthermore, Alghamdi, Lin, and Luo explored the integration of users' posting behavior clues with deep learning techniques, namely Convolutional Neural Networks (CNNs) and Bidirectional Gated Recurrent Units (BiGRUs), for the improvement of fake news detection on social media platforms [21]. Naik and Kumar emphasised the effectiveness of Long Short-Term Memory (LSTM) models in the automatic detection of fake news. They demonstrated the ability of these models to address the challenges presented by the rapid dissemination of misinformation on social media platforms [22]. Additionally, Monti and Sahoo [23] [24] emphasize the significance of considering social network structure and user behavior, in addition to content-based analysis. Monti's model, based on geometric deep learning, achieved an accuracy of 92.7% in detecting fake news on Twitter. In a further development, Kaliyar proposed a deep convolutional neural network (FND Net), [25] that attained a remarkable 98.36% accuracy in detecting fake news. Sedik [26] subsequently improved upon these results by proposing a deep learning-based system that combines concatenated and recurrent modalities, achieving an impressive 99.6% accuracy.

The recent advancements in deep learning have significantly enhanced the ability to detect fake news by integrating textual and contextual features. Mouratidis et al. [27] proposed a schema for textual inputs that are pairwise in nature, incorporating both the content of the news items in question and their context. The combination of content and context enables the model to more effectively capture the semantics and detect inconsistencies that are often indicative of fake news.

In a recent study, Bhatia et al. [28] introduced a deep neural network model that integrates multiple contextual features, including tweet content, publishing time, and social graph information. This multi-modal approach enables the model to consider a range of elements pertaining to the news, including the timing of publication and the social network of users sharing the content, in addition to the textual content itself. The combination of these diverse features enables the model to achieve an accuracy rate of 98.7% on the FakeNewsNet dataset. The incorporation of temporal and social graph features facilitates an enhanced comprehension of the broader context of news dissemination, thereby augmenting the model's capacity to identify fake news with greater reliability. This comprehensive approach underscores the significance of integrating diverse contextual data to enhance the efficacy of detection.

In the same context, Rajakumaran et al. [29] concentrated their attention on the possibility of early detection of fake news by examining propagation patterns and user interactions. The deep learning model employs contextual features, including user behavior and the propagation dynamics of news articles, to identify instances of fake news at the earliest possible stage of their dissemination. By monitoring early signals and interactions, their approach enables detection within hours of initial propagation, thereby providing a timely response to emerging misinformation. This capability is crucial for mitigating the impact of fake news and ensuring that accurate information prevails. The integration of contextual features related to propagation patterns enhances the model's sensitivity to early indicators of fake news, thereby demonstrating the importance of timely detection in combating misinformation.

In general, these deep learning techniques provide effective solutions for the automatic detection of fake news on Twitter, outperforming traditional classifiers and potentially complementing content-based approaches.

Transformer models, a subset of deep learning architectures, differ significantly from traditional deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). While CNNs excel at handling spatial data and capturing local dependencies, and RNNs are adept at processing sequential data by maintaining temporal dependencies, they both have limitations in capturing long-range dependencies efficiently [30]. Transformers, on the other hand, utilize a self-attention mechanism that allows them to consider the entire context of a sentence simultaneously, making them highly effective at capturing long-range dependencies and contextual relationships across a text c.

The combination of the self-attention mechanism with the parallel processing capabilities of transformers results in more efficient training and a greater suitability for large datasets than that observed with RNNs, which are sequential in nature [31].

Furthermore, transformers benefit greatly from transfer learning through pre-training on extensive corpora followed by fine-tuning on specific tasks, leading to significant performance improvements in natural language processing tasks [32] [33]. Notable examples of transformer models include BERT, GPT, and T5, which have set new benchmarks in the field by leveraging these advanced techniques [32] [34].

Varieties of approaches are currently under consideration, including the TSRI model, which combines the characteristics of the text, source, and user response in order to obtain accurate predictions [35].

Recent research has focused on transformer-based approaches for fake news detection, leveraging both news content and social contexts. These models aim to address challenges such as early detection and limited labeled data [36].The proposed frameworks utilize transformer architectures to learn representations from news articles and social media data [37].

The utilisation of transformative models, such as RoBERTa, has demonstrated promising results in the detection of fake news across diverse datasets, with RoBERTa exhibiting superior performance in terms of accuracy and training time compared to other models [38]. Similarly, Mina Schütz and colleagues [39] investigated the potential of pre-trained transformer models for the detection of fake news, using the FakeNewsNet dataset. The methodology entailed the application of transformer-based models to the text of news articles and a combination of the text of news articles and their titles, with an accuracy rate of up to 85%. In their study, Divyam Mehta et al., [40] examine the effectiveness of BERT. The authors fine-tune BERT on LIAR datasets and improve its performance by incorporating additional inputs such as human justification and metadata. Their methodology involves adapting BERT's deep contextual capabilities to classify news articles into both binary and six-label categories. The study shows that BERT significantly outperforms traditional models in binary classification, achieving 74% accuracy. While the improvement in the more complex six-label classification is modest, it remains important and highlights BERT's ability to capture and utilise nuanced textual and contextual information.

These advances underscore the significant potential of transformer-based models in enhancing the accuracy and efficiency of fake news detection on platforms like Twitter, pushing the boundaries of what is achievable in this critical area of research.

Table I provides a comprehensive overview of the different machine learning, deep learning and transformer models used for fake news detection.

This comprehensive analysis compares and contrasts traditional machine learning, deep learning, and transformer models to illustrate the advancements in the detection of fake news. The application of traditional machine learning techniques, including logistic regression, SVM and Naive Bayes, is contingent upon data pre-processing and feature extraction. However, these techniques often encounter difficulties in addressing complex relationships and handling large data sets. Deep learning models, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) units have advanced the field by automatically extracting complex patterns and processing large amounts of data, achieving significant results in terms of accuracy and user behavior analysis.

TABLE I. LITERATURE REVIEW FINDINGS SUMMARY

References	**Literature review findings summary**			
	Category	***Model***	***Dataset***	***Results***
[10], [11]	Machine Learning	SVM	Twitter data	99.36%
[12]		SVM	Twitter	93%
[13]		MNB, RF, SVC, LR	Twitter	79.05%.
[15], [16], [17]		Logistic Regression, SVM	Twitter	_
[18]		Passive-Aggressive Classifier	-	93%
[20], [21]	Deep Learning	CNN, BiGRU	Twitter	90%.
[23], [24]		Geometric Deep Learning	Twitter	92.7%
[25], [26]		FND Net, DNN	Twitter	99.6%.
[28]		DNN with Contextual Features	FakeNewsNet dataset	98.7%.
[41]	Transformer	BERT,	LIAR	74%
[39]		BERT	FakeNewsNet	85%

III. Proposed Multi-Model Embedding Approach

Before presenting the approach and methodology of the proposed fusion framework, it is imperative to define the problem in order to establish the objective of the proposed model.

A. Problem Definition

In the context of the detection of fake news on Twitter, the problem is defined as a supervised learning task with the objective of determining whether a specific tweet is true or fake. The task is formulated as a binary classification problem with the objective of accurately classifying tweets based on both the textual content and supplementary features derived through the utilisation of multiple transformer models.

Given a collection of tweets T= {T1, T2… Tn}, each tweet Ti consists of textual data. The corresponding labels L indicate the authenticity of each tweet, where:

- 0 indicates that the tweet is real.
- 1 indicates that the tweet is fake.

The objective is to model a prediction function F that takes as input a comprehensive feature vector **X(T)** derived from a fusion of multiple transformer models (BERT, ELECTRA, and XLNet) and additional handcrafted features. The function F predicts the label of the tweet, i.e., $F(T) \rightarrow \{0,1\}$, where:

- F(T)=0 if the tweet T is predicted to be real.
- F(T)=1if the tweet TTT is predicted to be fake.

In this fusion approach, the feature vector **X(T)** is composed of:

- Transformer-Based Embeddings:
  - CLS token embeddings from BERT and ELECTRA.
  - Mean-pooled embeddings from XLNet.
- Additional features:
  - Sentiment label and sentiment score.
  - Tweet length and word count.

The prediction function **F** is implemented as a machine learning classifier (e.g. Random Forest, XGBoost) that is trained on the fused feature vector X (T). The objective is to classify tweets as true or false using the diverse linguistic and contextual information provided by the fusion of multiple models and additional features.

The following section will provide a comprehensive explanation of the proposed approach.

B. Approach and Methodology

The objective of this article is to introduce the Multi-Model Embedding Approach to Fake News Detection (MT-FND), which represents a novel approach specifically designed for detecting fake news. It works by leveraging advanced natural language processing (NLP) techniques and integrating multiple transformer models—BERT, ELECTRA and XLNET.

The transformer models are selected for their robustness in NLP tasks, particularly in generating contextual embeddings that capture semantic nuances and contextual understanding, as demonstrated by prior research.

To operationalize the selected models within MT-FND, specific functions are employed for tokenizing text entries and extracting embeddings. For BERT and ELECTRA, embeddings are derived from the [CLS] token, while XLNet calculates embeddings by averaging all token embeddings. These embeddings serve as fundamental representations of the textual content. Additionally, technical features such as tweet length, word count, sentiment tags, and sentiment scores obtained through sentiment analysis are incorporated into the feature vector. Furthermore, MT-FND employs graph representation techniques using NetworkX in order to extend the scope of its analysis beyond that of textual embeddings. In this context, each tweet within the dataset is treated as a node within a graph, and is subsequently enriched with a number of attributes including labels, sentiment metrics, tweet length, word count, and sentiment scores. A placeholder function is employed to convert these attributes into preliminary embeddings, thereby establishing the foundation for representing articles within the graph structure.

This approach enables MT-FND to identify relationships between tweets, potentially revealing patterns that may not be discernible through text analysis alone.

The next step in the MT-FND approach involves using these enriched feature vectors to train two different classifiers: the Random Forest Classifier and the XGBoost Classifier.

The Random Forest Classifier is employed to utilise the combined embeddings derived from BERT, ELECTRA, and XLNet, in conjunction with supplementary features, to ascertain the veracity of a given tweet. This combination exploits the diverse strengths of each transformer model, thereby facilitating predictions that are more accurate. In contrast, the XGBoost Classifier is configured to assess the discrete embeddings from BERT and XLNet individually, thereby enabling it to capitalise on the distinctive capabilities of each model in formulating its predictions. The dual-classifier approach guarantees that MT-FND will benefit from both ensemble learning and the distinctive capabilities of each transformer model.

In summary, MT-FND improves the detection and classification of fake news by combining multiple transformer models (BERT, ELECTRA, and XLNet) with cutting-edge natural language processing techniques. It offers a comprehensive view of text content by fusing text elements with technical features like sentiment scores and tweet length. By utilizing Random Forest and XGBoost classifiers, MT-FND seeks to increase the precision of disinformation identification. Subsequent enhancements could involve utilizing NetworkX for graph representations, which could lead to better comprehension of relational data models via sophisticated graph neural network methods. This methodical approach uses complementary machine learning techniques and sophisticated transformers to tackle the challenges of detecting fake news. Fig. 1 provides an overview of the approach proposed and tested in this article.

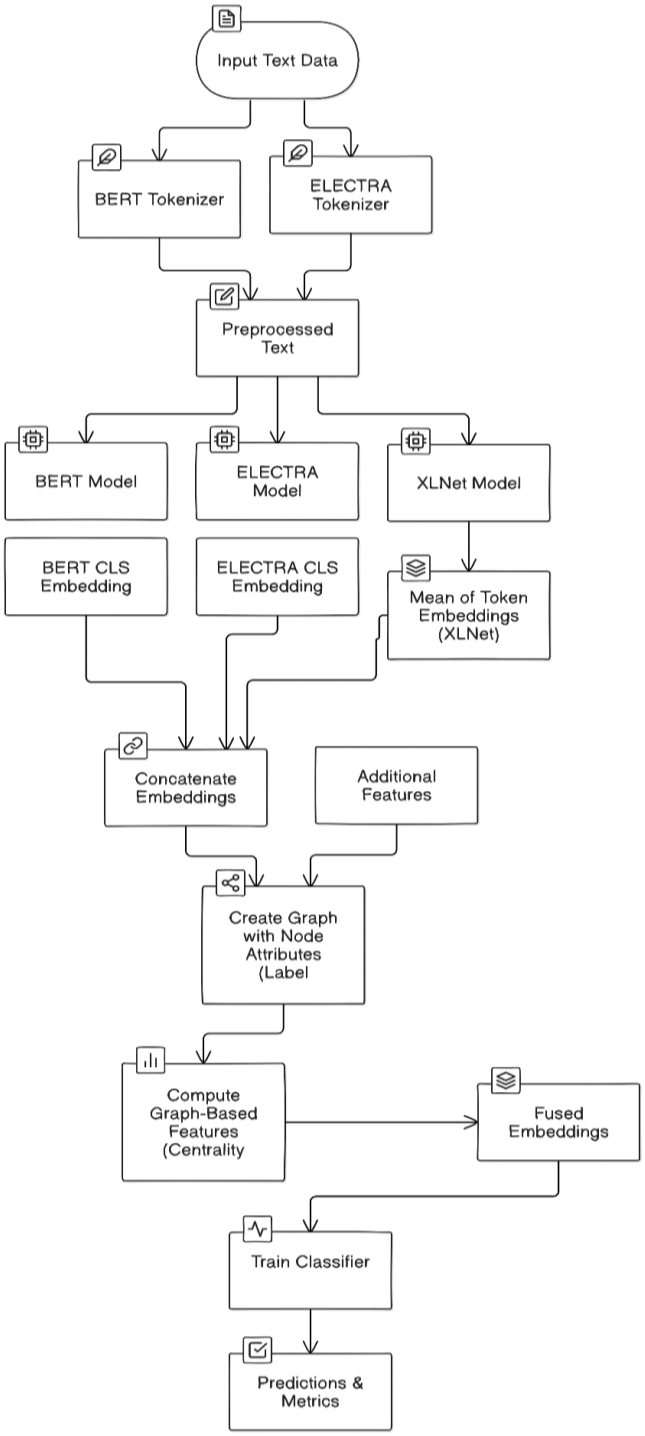

Fig. 1. Multi-model embedding approach to fake news detection (MT-FND).

IV. Experimental Set Up

In MT-FND approach, we opted for BERT, ELECTRA, and XLNet due to their established proficiency in natural language understanding tasks, each offering unique advantages such as contextual understanding, discriminative training efficiency, and bidirectional context capturing. By incorporating these diverse perspectives, we aimed to enhance the detection capabilities, effectively capturing the intricate nuances present in news articles, particularly within the constrained context of Twitter.

A. Dataset Description

For the purposes of this study, we utilized the FakeNewsNet dataset. The dataset is a comprehensive collection that was designed to help researchers study fake news detection and analysis. Two primary sources of data are included: BuzzFeed and PolitiFact. The dataset contains various features such as news content, social context, and spatiotemporal information. In the dataset, there are 11,510 news articles, with a proportionally balanced number of real and fake articles, with 5,755 articles labeled as real (label 1) and 5,755 articles labeled as fake (label 0). The balanced distribution of this dataset makes it possible for models trained on it to learn how to distinguish between real and fake news, which provides strong evaluation metrics and makes detection algorithms more reliable in real-world situations.

Fig. 2 shows the balance distribution of the dataset:

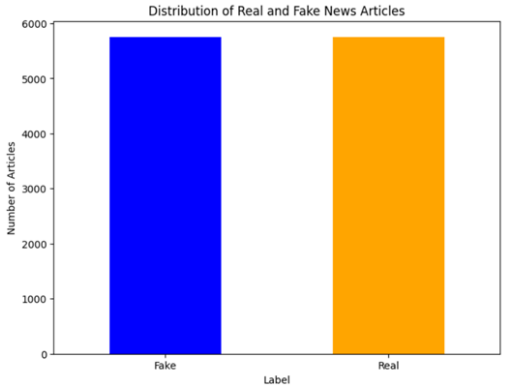

Fig. 2. Distribution of the balanced dataset.

B. Feature Engineering and Sentiment Analysis

The initial phase of feature engineering comprises the tokenization and embedding of news articles' titles, which is conducted through the utilisation of three distinct pre-trained transformer models. The models used are BERT, ELECTRA, and XLNet. Each model processes the text differently, thereby capturing various nuances of language. For example, BERT focuses on contextual relationships between words by employing bidirectional attention, while XLNet emphasize sequence order by considering permutations of input tokens. The output from each model is a high-dimensional vector representing the semantic content of the article's title. These embeddings serve as the foundational features for the classification model. In addition to the text embeddings, a sentiment analysis pipeline is utilised for the evaluation of the titles in question. This phase of the process classifies each ‘tweet’ as either "positive" or "negative" and assigns a corresponding sentiment score.

In order to represent the relationships between the tweets in a graph, the proposed approach involves constructing a graph using the NetworkX library. In this graph, each node represents a tweet, and edges are created based on cosine similarity between the text embeddings. Should the similarity between two articles exceed a specified threshold, an edge is formed, with the strength of this connection quantified by the similarity score.

Subsequently, centrality measures (such as degree centrality) and clustering coefficients are calculated for each node, thereby providing insights into the article's importance and its tendency to cluster with similar articles. These graph-based features are of paramount importance for understanding the broader context in which an article exists.

In the final stage, the extracted features are integrated into a unified representation for each article. The embeddings from BERT, ELECTRA, and XLNet are concatenated with sentiment analysis features and graph-based features such as centrality and clustering coefficients. This fusion creates a comprehensive feature set that captures both the textual and relational properties of each article.

C. Evaluation Criteria

The evaluation criteria used in this study are:

- Accuracy: The purpose of the accuracy metric is to ascertain the proportion of correctly predicted labels out of the total number of predictions. It is a simple yet effective metric for assessing the overall performance of the MT-FND model.

Metric: the accuracy metric is calculated during the testing phase by comparing the predicted labels (test_predictions) with the true labels (test_labels). The accuracy is computed using the accuracy_score function from the sklearn.metrics library, which divides the number of correct predictions by the total number of predictions. This metric provides a clear indication of the model's effectiveness in distinguishing between real and fake news articles.

- Precision: The purpose of the precision metric is to measure the accuracy of the positive predictions made by the MT-FND model. Specifically, it assesses the proportion of true positive predictions out of all positive predictions made by the model.

Metric: precision is calculated during the testing phase by comparing the predicted labels (test_predictions) with the true labels (test_labels). The precision is computed using the precision_score function from the sklearn.metrics library, which divides the number of true positive predictions by the total number of positive predictions. This metric is particularly important in scenarios where the cost of false positives is high, as it reflects the model's ability to avoid incorrect positive classifications.

- Recall: The purpose of the recall metric is to measure the model's ability to correctly identify all actual positive instances. It calculates the proportion of true positive predictions out of all actual positive cases.

Metric: recall is calculated during the testing phase by comparing the predicted labels (test_predictions) with the true labels (test_labels). The recall is computed using the recall_score function from the sklearn.metrics library, which divides the number of true positive predictions by the total number of actual positive instances. This metric is crucial in scenarios where missing a positive instance (false negatives) is costly, as it reflects the model's effectiveness in capturing all relevant instances.

- F1 Score: The purpose of the F1 score is to provide a balance between precision and recall, offering a single metric that accounts for both false positives and false negatives.

Metric: the F1 score is calculated during the testing phase by comparing the predicted labels (test_predictions) with the true labels (test_labels). The F1 score is computed using the f1_score function from the sklearn.metrics library, which combines precision and recall into a single metric.

V. Results and Discussions

A. Performance of MT-FND

In our evaluation of various classification models on a balanced dataset of real and false news articles, we introduced the Multi-Model Embedding Approach to Fake News Detection (MTFND), which integrates BERT, ELECTRA, and XLNet embeddings with graph features. This approach significantly enhances the accuracy and robustness of misinformation detection. The results clearly demonstrate the superiority of MTFND over models utilizing individual embeddings.

The performance of the proposed MTFND framework, which integrates fusion embeddings with graph features, demonstrates a strong enhancement in predictive capabilities across various classifiers. Among the models evaluated, the XGBoost classifier within the MTFND framework stands out, achieving an impressive accuracy of 87.28%. This model also shows a high precision of 85.56%, a recall of 89.53%, and an F1-score of 87.50%. These results highlight the ability of the MTFND framework to effectively capture and leverage complex features, providing a significant improvement over individual embedding approaches.

Table II presents a detailed comparison of model performance across various approaches, including the proposed MTFND framework and individual embeddings from BERT, XLNet, and ELECTRA.

TABLE II. Findings of the Experiments

Approach (Embedding type)	Model	Results by metrics type			
		Accuracy	*Precision*	*Recall*	*F1-score*
Proposed framework MTFND (Fusion embeddings + Graph Features)	RF	85.55	87.65	82.56	85.03
	SVM	84.97	84.09	86.05	85.06
	XGB	87.28	85.56	89.53	87.50
Individual BERT embeddings:	RF	84.97	84.09	85.06	84.97
	SVM	86.71	84.62	87.01	86.71
	XGB	84.97	84.09	85.06	84.97
Individual ELECTRA embeddings:	RF	79.19	82.89	73.26	77.78
	SVM	77.46	78.31	75.58	76.92
	XGB	81.50	83.75	77.91	80.72
Individual XLNET embeddings:	RF	78.03	79.27	75.58	77.38
	SVM	75.72	72.92	81.40	76.92
	XGB	78.61	77.53	80.23	78.86

When we evaluated models without the MTFND enhancements—relying solely on BERT embeddings or other contextual embeddings—the performance metrics were generally lower, highlighting the benefit of our approach.

For instance, the Random Forest classifier, when using only BERT embeddings without additional graph features, achieved an accuracy of 87.86%, a precision of 86.52%, and an F1 score of 88.00%. Although these figures are strong, they demonstrate that the inclusion of graph features in MTFND helps to provide a more nuanced understanding of the data, leading to predictions that are more accurate.

Similarly, the SVM classifier without the MTFND approach achieved an accuracy of 87.86%, a precision of 84.95%, and an F1 score of 88.27%. Again, while these results are respectable, they fall short of those achieved using the MTFND-enhanced models, particularly in terms of precision and F1 score, where the nuanced data representation provided by fusion embeddings and graph features shows its value.

The XGBoost classifier, even though robust, showed a noticeable drop in performance without the MTFND approach, recording an accuracy of 83.02%, a precision of 81.81%, and an F1 score of 79.27%. This further underscores the efficacy of the MTFND approach in enhancing the model's capability to detect and classify misinformation.

In conclusion, the MTFND approach significantly improves the performance of traditional machine learning models by enhancing their ability to analyze and interpret complex data patterns in news articles. The results clearly indicate that incorporating multimodal features, such as those in MTFND, is a powerful strategy for advancing the state of misinformation detection.

B. State-of-the-Art Comparison

Table III presents a comparative analysis of the proposed MTFND framework with several recent models from the literature. The MTFND framework achieved an accuracy of 87.28% using the XGB model, thereby demonstrating a significant improvement over the approaches described in study [39]. In particular, the BERT model proposed in study [39] achieved an accuracy of 85.0%. This comparison demonstrates that the MTFND framework, which integrates fusion embeddings (BERT, ELECTRA and XLNET) with graph features, exhibits superior performance in the detection of fake news in comparison to these transformer-based approaches. The XGB model within MTFND not only outperforms BERT and ALBERT in terms of accuracy but also demonstrates the efficacy of integrating diverse features and techniques. This suggests that the MTFND approach effectively harnesses both global text semantics and advanced features, resulting in a notable enhancement in performance. Further optimization and refinement of the MTFND framework may potentially yield even greater improvements.

TABLE III. COMPARATIVE MT-FND WITH OTHER WORKS

References	**Comparative MT-FND with other works**		
	Dataset	***Technique***	***Accuracy***
[39]	FakeNewsNet	BERT	85.0%
[40]	LIAR	BERT	74%
[13]	MNB, RF, SVC, LR	Twitter	79.05%.
Proposed framework MTFND	FakeNewsNet	Fusion embeddings (BERT,ELECTRA,XLNET)+ Graph Features with XGB classifier	87.28%

VI. CONCLUSION AND FUTURE WORK

This study presents a comprehensive approach to the detection of fake news, which integrates transformer-based embeddings with graph-based features. The integration of embeddings derived from BERT, ELECTRA, and XLNet with graph-based metrics has yielded a promising enhancement in performance, with notable improvements in accuracy, precision, recall, and F1-score values.

The fusion of transformer models' embeddings with additional features, such as sentiment scores, tweet length, and word count, provided a robust feature set that captures nuanced patterns in the data. By incorporating graph-based features like centrality and clustering coefficients, the model further enriched the representation of the data, enabling it to leverage the interrelationships between data points for improved classification. This hybrid approach outperforms traditional methods that rely solely on textual or structural features, underscoring the effectiveness of combining diverse sources of information.

By integrating the embeddings derived from three language models (BERT, ELECTRA, and XLNet) with graph-based features, the framework attains an exceptional accuracy of 87.28% on the FakeNewsNet dataset. This fusion approach exploits the distinctive capabilities of each model. The combination of BERT's contextual embeddings, ELECTRA's efficient training, and XLNet's permutation-based learning allows for the capture of a comprehensive representation of the text. Furthermore, the incorporation of graph features derived from cosine similarity and centrality measures enhances the feature set by capturing relational dynamics and structural patterns among articles.

This comprehensive approach not only enhances the model's capacity to discern subtle distinctions between authentic and fabricated news items but also improves its resilience and generalisability. The fusion of diverse embeddings with graph-based insights represents a significant advancement in the field of fake news detection, offering a more nuanced and effective solution.

Further research could enhanced by the incorporation of additional graph-based metrics, such as those pertaining to community detection and influence propagation. These advanced graph features have the potential to capture more intricate relationships and patterns within the data, thereby enhancing the model's ability to understand and detect nuanced misinformation. Furthermore, the investigation of novel or domain-specific transformer models, such as GPT-3, may facilitate enhancements in feature extraction and classification accuracy. By leveraging the latest advancements in model architecture and integrating these with graph-based insights, future models could achieve enhanced performance in the detection of fake news.

Finally, the expansion of the dataset through augmentation and enrichment represents another promising avenue of enquiry. The incorporation of diverse sources and the generation of synthetic data could improve the model's generalisation and robustness, addressing class imbalances and simulating various news scenarios. Furthermore, the development of real-time detection capabilities and the integration of the system with social media platforms would facilitate the timely and effective detection of fake news.

REFERENCES

[1] Lahrou, Y. (2019). Automatic detection of fake news on online platforms: A survey. In Proceedings of the 1st International Conference on Smart Systems and Data Science (ICSSD).

[2] Shu, K., Wang, S., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

[4] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[5] Yang, Z., Dai, Z., Yang, Y., Cohen, W. W., & Salakhutdinov, R. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In Proceedings of NeurIPS 2019: Advances in Neural Information Processing Systems.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., & Clark, K. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[7] Lahrou, Y. (2021). Automatic detection of fake news on Twitter by using a new feature: User credibility. In Proceedings of the 5th International Conference on Big Data Cloud and Internet of Things (BDIoT).

[8] Yang, Z., Dai, Z., Yang, Y., Cohen, W. W., & Salakhutdinov, R. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In Proceedings of NeurIPS 2019.

[9] Liu, Y., Ott, M., Goyal, N., Du, J., & Clark, K. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[10] Shabbir, A. S., & Khan, M. I. (2023). Fake Twitter followers detection using machine learning approach. In Proceedings of the International Conference on Business Analytics for Technology and Security (ICBATS).

[11] Hisham, M. (2023). An innovative approach for fake news detection using machine learning. University Research Journal of Engineering and Technology, 13.

[12] Raja, L. R. (2022). Fake news detection on social networks using machine learning techniques. Materials Today: Proceedings.

[13] Srinivas, J. S., & Pal, R. J. (2021). Automatic fake news detector in social media using machine learning and natural language processing approaches. In Proceedings of Smart Computing Techniques.

[14] Yadav, S. (2023). Machine learning based approach to disinformation detection using Twitter data. In Proceedings of the International Conference for Advancement in Technology (ICONAT).

[15] Zhou, X. (2018). Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315.

[16] Shu, K., Wang, S., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. arXiv preprint arXiv:1708.01967.

[17] Shetty, M. S., & Swamy, A. S. (2023). Automated method for fake news detection using machine learning. In Proceedings of the International Conference on Network, Multimedia and Information Technology (NMITCON).

[18] Bhogi, A. D., & Nair, S. K. (2023). Machine learning for fake news detection on social media text. In Proceedings of the International Conference on Advances in Computation, Communication and Information Technology.

[19] Nguyen, V. S., & Do, T. H. (2020). FND: A framework for fake news detection on social media platforms. Computers in Human Behavior.

[20] Bhatia, T. M., & Kumar, A. (2023). Detecting fake news sources on Twitter using deep neural network. In Proceedings of the 11th International Conference on Information and Education Technology (ICIET) (pp. 508-512).

[21] Alghamdi, J. L., & Aljohani, S. (2023). Does context matter? Effective deep learning approaches to curb fake news dissemination on social media. Applied Sciences: Multidisciplinary Digital Publishing Institute, 13.

[22] Patel, U., & Gupta, P. (2023). Fake news detection using neural network. In Proceedings of the IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS).

[23] Monti, F., & Defferrard, M. (2019). Fake news detection on social media using geometric deep learning. Computer Science.

[24] Sahoo, S. R., & Bhunia, A. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. Applied Soft Computing.

[25] Kaliyar, R. K., & Sahu, M. (2020). FNDNet – A deep convolutional neural network for fake news detection. Cognitive Systems Research.

[26] Sedik, A. M., & Alshamrani, S. (2022). Deep fake news detection system based on concatenated and recurrent modalities. Expert Systems with Applications.

[27] Mouratidis, D., & Papatheodorou, C. (2021). Deep learning for fake news detection in a pairwise textual input schema. De Computis.

[28] Bhatia, T. M., & Sharma, A. (2023). Detecting fake news sources on Twitter using deep neural network. In Proceedings of the International Conference on Innovation Engineering and Technology.

[29] Rajakumaran, K. A. R., & Kumar, K. (2021). Fake news detection in Twitter datasets using deep learning techniques. Computer Science.

[30] LeCun, Y., & Bengio, Y. (2015). Deep learning. Nature, 521(7553), 436-444.

[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., & Jones, L. (2023). Attention is all you need. arXiv preprint arXiv:1706.03762.

[32] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[33] Radford, A., Narasimhan, K., & Salimans, T. (2019). Language models are unsupervised multitask learners. In Proceedings of the 2019 OpenAI.

[34] Raffel, C., Shinn, C., & Roberts, A. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

[35] Luo, Y., Wu, X., & Yang, Z. (2023). Social media fake news detection algorithm based on multiple feature groups. In Proceedings of the IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA).

[36] Raza, S. (2022). Fake news detection based on news content and social contexts: A transformer-based approach. Computer Science.

[37] Do, T. H., & Tran, T. D. (2021). Context-aware deep Markov random fields for fake news detection. Computer Science.

[38] Babu, N. R., & Kumar, A. (2023). Performance evaluation of transformer-based NLP models on fake news detection datasets. Engineering, Electrical and Computer.

[39] Schütz, M., & Mehta, A. (2020). Automatic fake news detection with pre-trained transformer models. In Proceedings of the ICPR Workshops.

[40] Mehta, D., & Patel, R. (2021). A transformer-based architecture for fake news classification. Social Network Analysis and Mining.

[41] Kalkatawi, M. M., & Alotaibi, M. (2023). The detection of fake news in Arabic tweets using deep learning. Applied Science.