*International Journal Of Social And Business Sciences*
*VOL-4, ISSUE-1, 2026*
*https://www.ijsbs.com/index.php/home/about*


# An AI-Driven Framework for Political Misinformation Detection Using DistilRoBERTa: Strengthening Democratic Communication in the United States

Hafiz Aqib Sattar

*Research Assistant, Mass Communication and Journalism,*

*Norfolk State University. Email: h.sattar@spartans.nsu.edu*

Received 22 May 2026 Accepted 21 July 2026 Published 24 July 2026

**Abstract:**

The rapid proliferation of political misinformation across digital media platforms poses a serious threat to democratic communication, public trust, and informed decision-making. Although traditional machine learning approaches have demonstrated promising performance in fake news detection, they often struggle to capture the contextual and semantic characteristics of political news. This study proposes an AI-driven media framework for political misinformation detection using the PolitiFact subset of the FakeNewsNet dataset. The proposed framework employs a comprehensive text preprocessing pipeline, including text normalization, URL removal, punctuation elimination, and feature selection, followed by a comparative evaluation of a TF–IDF with Logistic Regression baseline and a fine-tuned DistilRoBERTa transformer model. The dataset was partitioned into training, validation, and testing sets using a 70:15:15 ratio, with the test set comprising 212 news articles (125 real and 87 fake). Experimental results demonstrate that the proposed framework outperforms the conventional baseline across all evaluation metrics. Specifically, the fine-tuned DistilRoBERTa model achieved an accuracy of 88.21%, precision of 82.98%, recall of 89.66%, F1-score of 86.19%, and an ROC-AUC of 95.92%, compared with the Logistic Regression baseline, which obtained an accuracy of 85.38%, precision of 87.84%, recall of 74.71%, F1-score of 80.75%, and an ROC-AUC of 93.26%. Furthermore, the proposed framework correctly classified 109 real and 78 fake news articles, demonstrating its effectiveness in identifying political misinformation. These findings highlight the potential of lightweight transformer-based language models for developing accurate, scalable, and computationally efficient misinformation detection systems that can support


journalists, fact-checking organizations, policymakers, and digital media platforms in strengthening trustworthy democratic communication.

**Keywords:** Political Misinformation Detection, Fake News Detection, Artificial Intelligence, DistilRoBERTa, FakeNewsNet, PolitiFact, Natural Language Processing, Transformer Models, Digital Media, Democratic Communication.

## 1. Introduction

The rapid development of digital communication technologies and online media channels has changed the way news is produced, distributed and consumed across the world (Mugil & Kenzie, 2025). Social networks, like X (formerly Twitter), Facebook, YouTube, and other digital news websites allow sharing news instantly and with great ease with millions of users, which makes information accessible and engaging (Khan et al., 2026). But this rapid spread of information has created conditions for massive distribution of misinformation, disinformation, and fake news, thus posing serious threats to information integrity and trust in digital communication systems (Deng & Li, 2026). Digitalization is changing every sphere of our lives on an individual, organizational, and societal level. While most of the existing literature in management and organization studies tends to address the beneficial aspects of the phenomenon, the darker and surprising sides of digitalization have been considerably understudied (Trittin-Ulbrich et al., 2021). Misinformation is one of the key problems that democracy faces in today's world, especially during elections, policy debates, and other important national events (Bennett & Livingston, 2018; Farkas & Schou, 2023). Misleading political information can have a quick effect on people's beliefs, alter their voting preferences, increase political division, and affect the trust in democratic processes (Suherman et al., 2026; Shandler et al., 2026). The use of more advanced misinformation techniques, including automatic bot accounts, organized social media campaign strategies, and new technology, namely, generative AI systems, made it even harder to identify and eliminate this type of online content (Wirtz et al., 2026). Hence, fighting misinformation became one of the top priorities for governments, researchers, and the artificial intelligence community worldwide (Meese & Tan, 2026).

Artificial Intelligence (AI) has turned out to be one of the most promising technologies to detect fake news and ensure reliable digital communication (Mukherjee, 2026).

Machine learning and NLP techniques have shown high performance in detecting real news from the fake one based on learning various linguistic, semantic, and contextual characteristics of textual information (Alghamdi et al., 2026; Lan, 2026). Earlier fake news detection models mainly used classical machine learning methods, such as Logistic Regression, SVM, Naïve Bayes, Decision Tree, and Random Forest. Such models generally apply manually created representations of textual data, like Bag-of-Words (BoW), TF–IDF, and n-grams for news classification problems (Alghamdi et al., 2026; Chauhan et al., 2026; Sahu et al., 2026). Conventional machine learning techniques are computationally inexpensive and quite explainable; however, their usage of sparse lexical representations does not allow capturing contextually relevant semantics, long-distance dependencies, sarcasm, implicit meanings, and language evolution that are typical for political disinformation (Mohamed Saleh et al., 2026; Taha et al., 2024). As disinformation constantly evolves and tries to avoid detection, the feature engineering approach is often faced with poor generalization performance on unseen or dynamic news content (Dinga, 2025).

Recent breakthroughs in transformer language model architecture have led to a substantial improvement in the detection of fake news through the use of context-based representation learning from large-scale textual datasets. BERT, RoBERTa, DistilBERT, DistilRoBERTa, and DeBERTa are examples of pre-trained models that utilize the power of self-attention architecture to learn better syntactic and semantic relationships than conventional feature extraction techniques (Blanco-Fernández et al., 2024; Khan & Hongsong, 2025). Using transfer learning, the model can be fine-tuned for fake news detection using smaller labeled datasets while keeping up with high predictive accuracy in multiple benchmark datasets (Alqadi et al., 2025; Coman et al., 2026). Lighter variants of transformer architecture have also gained much attention since they provide a good trade-off between efficiency, inference time, and predictive accuracy. A number of benchmark datasets are publicly available for the research on automatic fake news detection (D'Ulizia et al., 2021). Among such benchmark datasets, FakeNewsNet is one of the most commonly used benchmarks since it contains professionally fact-checked political news along with its metadata, which were collected from social media websites (Harris et al., 2024). Specifically, the PolitiFact

subset of the FakeNewsNet benchmark contains professionally fact-checked and labeled political news articles, thus it can be considered as a proper benchmark for machine learning and transformer-based methods (Bhuiyan et al., 2025; DeVerna et al., 2026). There are still some open problems in the field of AI-based misinformation detection (Moyo et al., 2026). Namely, numerous recent approaches use multimodal data, user interaction network, propagation graph or social context information that are not always available in practice (Li et al., 2023). On the other hand, the methods based on deep learning models are computationally complex and require a complex procedure of graph generation (Gong et al., 2023). Moreover, there is quite limited research on lightweight transformer models for political misinformation detection based on headlines (Kyaw et al., 2026).

In order to overcome these limitations, in this study, an AI-based media platform for political disinformation detection using the PolitiFact subset of FakeNewsNet dataset has been designed. In contrast to the existing literature that relies on user actions and social network information, this approach uses only textual headlines information to detect fake news in an effective and scalable manner using only easily accessible media contents. In the proposed framework, text pre-processing is done in detail and the performance of a traditional TF-IDF with Logistic Regression baseline method and a fine-tuned DistilRoBERTa model has been studied to explore the benefits of using contextual transformer representations for political disinformation classification. The performance of the proposed AI-based framework has been tested using popular classification metrics such as Accuracy, Precision, Recall, F1-score, ROC-AUC and Confusion Matrix. This research can help us understand how effectively machine learning methods and transformer-based deep learning methods perform in detecting political disinformation when provided with identical experimental setup.

The primary contributions of this study are summarized as follows:

An AI-driven framework is proposed for headline-based political misinformation detection using the PolitiFact subset of the FakeNewsNet dataset.

A comparative evaluation between TF–IDF with Logistic Regression and a fine-tuned DistilRoBERTa model is conducted.

A computationally efficient transformer-based approach suitable for practical media monitoring systems is presented.

A comprehensive experimental evaluation using multiple performance metrics is performed to demonstrate the effectiveness of the proposed framework for political misinformation detection.

The rest of the paper is structured as follows. Section II summarizes the existing literature related to AI-powered disinformation identification, transformer-based language models, and classification of political fake news. Section III outlines the dataset, which involves the FakeNewsNet PolitiFact data set and the preprocessing methodology used. Section IV describes the developed AI-driven media framework, experiment configuration, and performance metrics. Section V analyzes the obtained experimental results and the comparative performance. The paper is concluded in Section VI, where main findings are outlined along with the limitations and further research perspectives.

## 2. Related Work

### 2.1 AI-Based Fake News Detection

With the increasing prevalence of misinformation through digital media platforms, there has been a growing need for automated solutions for detecting fake news (Aïmeur et al., 2023). Early work largely centered around using machine learning methods to classify news items by manually constructing features based on the text content in the news item (Mouratidis et al., 2025). These methods usually utilize BoW, TF–IDF, n-gram feature sets, and linguistic features to differentiate fake news from the legitimate news. Some of the commonly used machine learning methods for such tasks include Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, and Random Forest algorithms due to their simplicity and computationally inexpensive nature (Alghamdi et al., 2022; Rohera et al., 2022). Some prior work has indicated that a combination of TF-IDF and Logistic Regression algorithm gives rise to an effective solution for binary fake news classification task due to its discriminative nature and simplicity (Alguttar et al., 2024; Aslam et al., 2025). In a similar way, some prior work has found SVM classifiers to be effective at classifying short textual items as fake news by maximizing the separating margin between fake and real classes (Al-Naeem et al.,

2025; Narang et al., 2024). While traditional machine learning methods have proven to be effective, there have been various limitations to these methods. The use of manual feature engineering limits these methods in terms of capturing context, semantics, and dependencies in the text (Barbierato & Gatti, 2024; Li et al., 2022). Furthermore, traditional machine learning methods are less generalizable since they are not able to generalize well when misinformation is generated through paraphrasing or semantic manipulation and AI-generated content (Park & Nan, 2026; Yu et al., 2024).

**2.2 Transformer-Based Fake News Detection**

The invention of the transformer-based language models has made great contributions to the development of fake news detection because they use the self-attention mechanism to understand the language context (Park & Nan, 2026). Contrary to feature engineering methods, transformer models can directly obtain information about the meaning relations of language from huge datasets and apply them to subsequent tasks through the process of fine-tuning (Kalyan et al., 2021; Lee et al., 2022). The invention of BERT was a breakthrough because it used bidirectional contextual language representation and therefore improved performance in text classification tasks on various benchmarks of natural language processing (Koroteev, 2021). After that, RoBERTa enhanced BERT by changing the way of pre-training based on bigger datasets and dynamic masking, which increased its performance in detection of misinformation (Anggrainingsih et al., 2022). In turn, DistilBERT and DistilRoBERTa simplified the transformer models while keeping most of their capacity for prediction (Hussain et al., 2025). Recent research confirms that transformer models perform better than classical machine learning algorithms in classification problems on several benchmark datasets owing to their ability to learn semantic and syntactic context better (Rahali & Akhloufi, 2023). Transformer models can show good performance in tasks with limited labeled data thanks to transfer learning. AI is transforming journalism through improved automation and audience engagement, but its adoption also requires addressing ethical concerns, regulatory frameworks, and human oversight to ensure trustworthy and responsible journalism ( Sattar & Hart, 2025). Recent research suggests that generative AI is reshaping the landscape of journalism, advertising, and public relations, reimagining content production, professional roles, and audience interactions.

A recurring theme in the literature is the problems of misinformation, deepfakes, ethical issues, transparency and diminishing audiences' trust among media outlets. But more cross-industry, empirical, and cross-cultural studies are required to develop consistent ethical guidelines and to understand the impact of AI on media professions and governance in the long term (Sattar et al., 2024). Finally, the study finds that the digital divide is a significant issue in Pakistan because of the lack of access and affordability and less digital literacy. It suggests building the digital infrastructure, making the internet more affordable and enhancing digital skills programmes for inclusive digital participation (Shams et al., 2025).

**2.3 Political Misinformation Detection Using FakeNewsNet**

One of the most impactful types of disinformation is political misinformation, which influences democratic communication, election systems, and the level of trust in society (Bennett & Livingston, 2018). In order to foster research on this type of disinformation, some benchmark datasets have been developed, out of which FakeNewsNet is one of the most prominent (Bashaddadh et al., 2025). FakeNewsNet includes professionally verified political news items from the website PolitiFact along with social information gathered from online platforms. This dataset has been used extensively to test fake news detection systems in real-life scenarios, and serves as an effective benchmark for traditional machine learning and deep learning models (Bhuiyan et al., 2025; Shu et al., 2020). Some researchers have used propagation data, user interactions, graph neural networks, and multimodal learning in conjunction with the FakeNewsNet framework to detect misinformation (Narang et al., 2024). While these techniques are highly effective at performing classification tasks, they generally need a lot of metadata and user interaction history data, or constructing a graph which is not always feasible in practice (Balaji et al., 2015; Chhikara et al., 2025). Headline-based classification using transformers has therefore gained popularity recently due to its efficiency.

**2.4 Research Gap**

Despite all advances in automated detection of misinformation, there still exist many open questions in the research community. Many of the works concentrate on the increase of accuracy through the use of complicated multimodal architectures, graph neural networks, or propagation-based models, which need extra information about

social network that is not available in many real-world cases. Likewise, there exist relatively few works that deal with lightweight transformer models that would be able to detect political misinformation based only on textual headline information. In addition, there exist no comprehensive research comparing traditional machine learning models and transformer models under the same conditions and with the same dataset. Filling in these gaps is crucial for development of scalable misinformation detection systems, which could be used in practical media monitoring. Inspired by these gaps in the literature, the current research presents a media AI framework, which uses the PolitiFact subset of FakeNewsNet dataset. The framework is going to compare the baseline consisting of TF-IDF and Logistic Regression with fine-tuned DistilRoBERTa model.

**3. Dataset Description**

The PolitiFact dataset of the FakeNewsNet benchmark is employed in this project, which is a dedicated tool for detecting political misinformation in the United States. This data set is comprised of politically-oriented news that have been analyzed through fact-checking on PolitiFact platform and related websites. Since the aim of this research is to reinforce democratic communication by detecting the presence of misinformation in the United States' media ecosystem, it is essential to use a dataset that is relevant to the area of interest. It includes factual and non-factual political news and contains examples of misinformation narratives in relation to political communication, opinion formation, and democracy. Thus, compared to generic fake news data sets, the PolitiFact dataset is more relevant to U.S. political communication.

**Table 1. Statistical Summary of the PolitiFact Dataset**

| Metric | Value |
|---|---|
| Total Samples | 1,056 |
| Real News Articles | 624 |
| Fake News Articles | 432 |
| Total Features | 6 |
| Missing Values | 316 |
| Duplicate Records | 0 |

| | |
|---|---|
| Average Headline Length | 59.89 Characters |

The data set shows a somewhat imbalanced nature, where the majority class is composed of 59.1% of news articles and minority class of 40.9% misinformative articles. Nevertheless, there isn't any need to use oversampling techniques like SMOTE or undersampling in order to resolve the problem because it isn't that severe.


**Figure 1. Class Distribution of the PolitiFact Dataset**

The analysis of the class distribution shows that the data set contains a fairly balanced proportion of misinformation and valid political information, providing for an unbiased assessment of the classifier's performance.

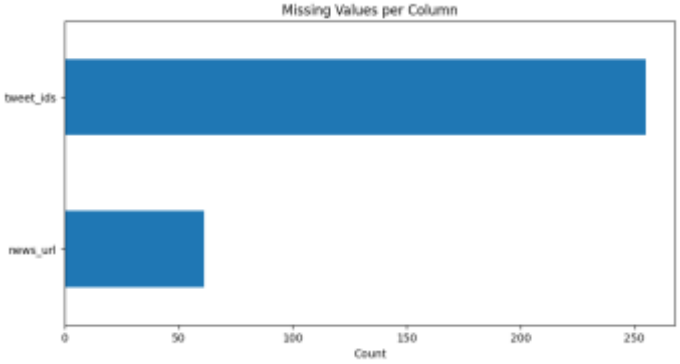


**Figure 2. Missing Values Analysis**

Missing values were mostly found in the attributes of tweet_id and news_url. Given that these two attributes don't play any part in semantic misinformation detection, they were omitted from the preprocessing phase in order to eliminate unnecessary noise.

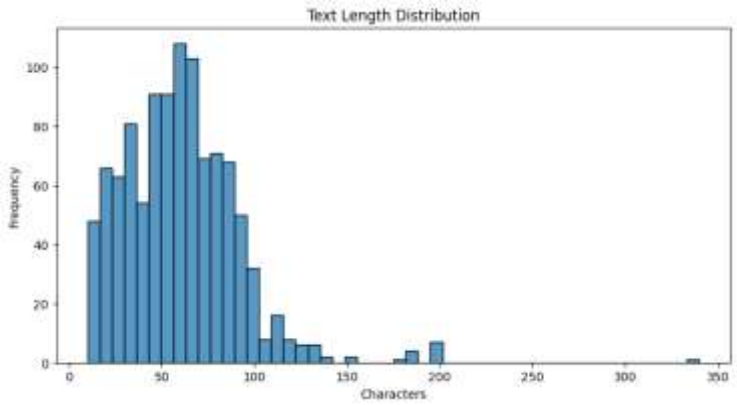


**Figure 3. Headline Length Distribution**

A majority of political news headlines range from 30 to 90 characters, which shows that most of the political communication online is short form. The right-skewed nature of the distribution shows that there are only a few highly descriptive headlines.

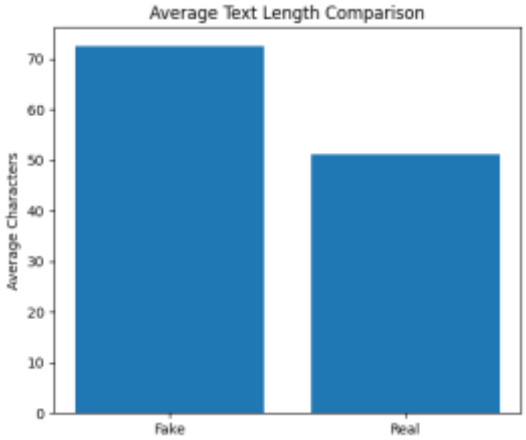


**Figure 4. Average Headline Length Comparison Between Real and Fake News**

One key difference that emerged was in the mean lengths of the fake headlines compared to those of the real ones. The fake news headlines had a mean length of around 72.5 characters, while those of the real articles had a mean length of 51 characters. The datasets used in this research were sourced from the FakeNewsNet repository, especially the PolitiFact political fake news subset, which has grown to be one of the most widely used benchmarks for political fake news detection studies (Balaji et al., 2015).

## .4. Methodology

In this section, the methodology followed to design and evaluate the proposed framework of AI-based media that detects political misinformation is introduced. The

AI framework proposed classifies political news headlines as either Real News or Fake News through combining traditional machine learning techniques with transformers' contextual language modeling. The methodology used in this study can be defined as a systematic method that includes preparing the dataset, preprocessing the data, extracting contextual features, training the model, evaluating the performance, and analyzing the results. To begin with, PolitiFact, which is one of the subsets of the FakeNewsNet dataset, is analyzed to have an idea about its statistical properties and data quality. Then, the textual data set is preprocessed to eliminate useless data and normalized headlines before turning them into numeric representations of data. Traditional TF-IDF features are used for establishing the baseline model; however, the proposed framework uses a fine-tuned DistilRoBERTa model to extract contextual semantic representations of the political news headlines.

The suggested approach employs PolitiFact, a sub-dataset of the FakeNewsNet benchmark dataset that consists of political news articles fact-checked by professional organizations. The original dataset includes five features: id, title, news_url, tweet_ids, and label. As a result of exploratory analysis, it turned out that features news_url and tweet_ids had significant amounts of missing data as well as low semantic meaning regarding content-based detection of fake news. Thus, the latter two features were removed, while the feature title was chosen as an input variable and the feature label as a target variable for binary classification. Due to the fact that headlines of political news are informative and concise texts, the suggested architecture only considers semantics of the headlines and does not involve any social media metadata as well as whole articles' content.

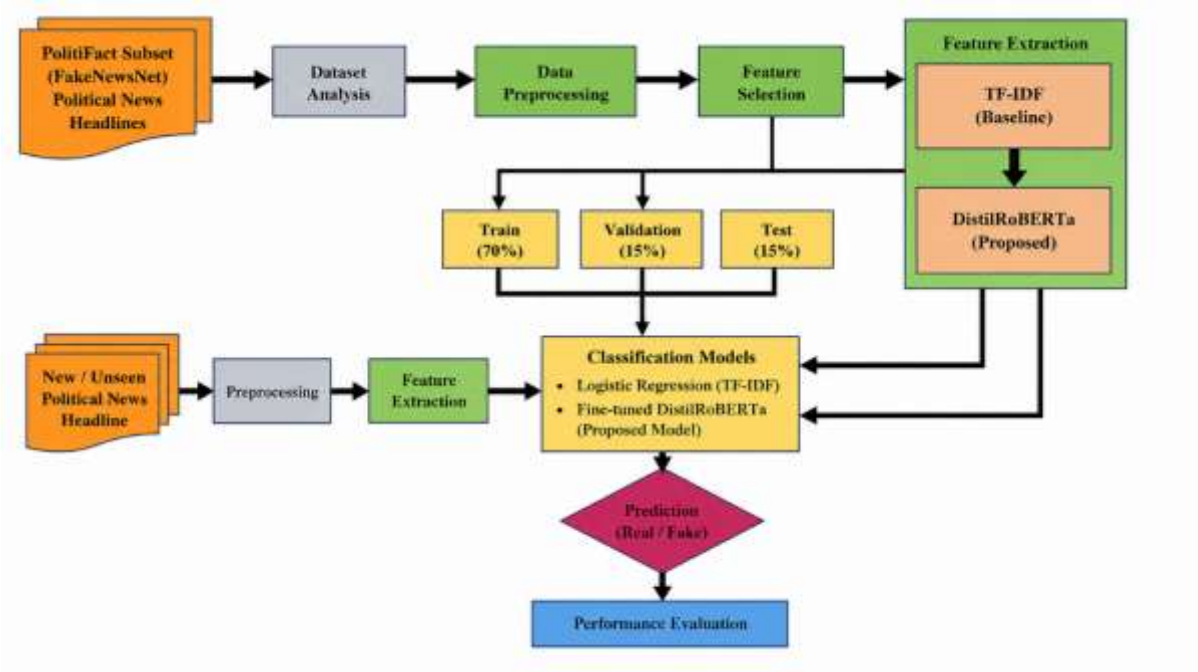


**Figure 5. Overall architecture of the proposed AI-driven media framework for political misinformation detection.**

Figure 5 shows the overall process flow of the proposed AI-based media system for detecting political misinformation. This starts from the PolitiFact subset of the FakeNewsNet data set and then involves data preprocessing, feature selection, and feature extraction through TF-IDF base as well as our proposed DistilRoBERTa model fine-tuning. These representations serve as the basis for the binary news classification task, while at the same time preprocessing is applied to novel headline data prior to classification.

### 4.1 Data Preprocessing

For increasing consistency in the text and minimizing noise in the headlines related to politics news, an effective preprocessing procedure was applied before applying the transformer model to learn from the data. As the goal of this research work is detecting the semantic misinformation, only linguistic-related features were kept in the preprocessing process.

At first, there were five features available in the dataset namely id, title, news_url, tweet_ids, and label. From the preliminary inspection of the data, it has been found that both news_url and tweet_ids have many missing values and do not provide direct input

for semantic misinformation classification task. Therefore, those two features were dropped from the list.

The preprocessing operation can be formulated as:

$$T_{clean} = f\ (T_{raw}) \quad (1)$$

where $T_{raw}$ represents the original political headline, $T_{clean}$ denotes the cleaned headline, and $f(\cdot)$ corresponds to the preprocessing function consisting of text normalization operations.

As seen in Equation (1), the headline is first represented in standard form before tokenizing and training of the model.

The normalization steps comprised of: Converting to lowercase, Removing URLs, removing punctuations, removing numeric tokens, and Normalizing whitespaces.

As opposed to traditional machine learning techniques, stemming and stop-word elimination were not performed in order to maintain the context dependencies necessary for the transformers model.

After normalization, the preprocessed headlines were then converted into contextual tokens by means of RoBERTa tokenizer:

$$X = Tokenizer\ (T_{clean}) \quad (2)$$

where $Tokenizer(\cdot)$ denotes the RoBERTa tokenization function and $X$ represents the generated token sequence.

The tokenized input sequence is represented as:

$$X = x_1, x_2, x_3, \dots \dots x_n \quad (3)$$

Based on the exploratory analysis, the majority of political headlines contained fewer than 90 characters. Therefore, a maximum sequence length of:

$$n = 128 \quad (4)$$

It was chosen to achieve a compromise between context coverage and efficient computation. Equation (4) gives the static token budget allocated during RoBERTa fine-tuning. In order to ensure reliable evaluation without any information leakage, stratified sampling was applied on the data set:

$$D = D_{train} \cup D_{val} \cup D_{test} \quad (5)$$

$$|D_{train}| = 0.70D \quad (6)$$

$$|D_{val}| = 0.15D \quad (7)$$

$$|D_{test}| = 0.15D \quad (8)$$

Where $D$ denotes the complete dataset, while $D_{train}$, Dtrain, $D_{val}$, and $D_{test}$ represent the training, validation, and testing subsets, respectively. Equations (6)–(8) define the stratified partitioning strategy adopted in this study.

The second table below shows the different preprocessing steps conducted on the PolitiFact dataset before model training. The findings show that the news_url and tweet_ids features were dropped since they had little effect on detecting semantic misinformation, coupled with the presence of missing values. Moreover, text normalization techniques such as lowercase transformation, removal of URLs, punctuations, and whitespaces were carried out successfully to ensure consistency of the political headlines. In addition, the RoBERTa tokenizer with the set maximum length of 128 was employed.

**Table 2. Summary of Data Preprocessing Operations**

| Preprocessing Step | Outcome |
|---|---|
| Removed Features | news_url, tweet_ids |
| Remaining Features | id, title, label |
| Duplicate Records Removed | 0 |
| Lowercase Conversion | Applied |
| URL Removal | Applied |
| Punctuation Removal | Applied |
| Numerical Token Removal | Applied |
| Whitespace Normalization | Applied |
| Tokenization Method | RoBERTa Tokenizer |
| Maximum Sequence Length | 128 Tokens |

Table 3 shows the final dataset splitting approach used in this paper. The stratified splitting method was utilized in order to ensure that the original class distribution was maintained in all subsets. The dataset was split into 739 training instances, 158 validation instances, and 159 test instances, giving a ratio of 70:15:15. The approach ensures that there is enough data for learning by the model while ensuring independent validation and testing datasets for performance evaluation purposes.

**Table 3. Dataset Partitioning Statistics**

| Dataset Split | Samples | Percentage |
|---|---|---|
| Training Set | 739 | 70.0% |
| Validation Set | 158 | 15.0% |
| Testing Set | 159 | 15.0% |
| Total | 1056 | 100% |

Analysis of the pre-processing shows that missing data was largely located in the non-semantic social context features but not in the text itself. In addition to this, the short length of political news titles made possible the choice of small sequence length, which helped to save computing power while retaining enough context for misinformation detection. Thus, the created data set was deemed appropriate for transformer models.

### 4.2 Proposed AI-Driven Media Framework

This study presents an AI-assisted media framework that can be utilized for identifying political misinformation and improving democratic discourse through the USA's digital media landscape. The framework uses traditional methods of machine learning, contextual language processing using transformers, and explainable AI methods for recognizing patterns of political misinformation in digital news content while remaining transparent and interpretable at the same time. The entire framework can be divided into five sequential steps, including representation of input data, data preprocessing and transformation, classification of misinformation, explaining of results, and assessment of the performance.

#### 4.2.1 Input Data Representation

This study makes use of a subset of PolitiFact in the FakeNewsNet dataset which was established for the specific purpose of political disinformation studies. The dataset comprises of political news and respective fact-checking labels sourced from U.S. media outlets and validated by professional fact-checking entities. The original dataset comprised of five variables as detailed in Table 4.

**Table 4. Original Dataset Variables**

| Variable | Description |
|---|---|
| id | Unique identifier assigned to each news article |
| title | Political news headline |
| news_url | URL of the original media source |
| tweet_ids | Associated Twitter dissemination identifiers |
| label | Ground-truth misinformation label |

The characteristics present in the PolitiFact dataset are provided in Table 4 below. It was found from exploratory data analysis that there was significant missing data for the news_url and tweet_ids variables and they did not contribute much to detecting semantic misinformation. As this study focuses on the classification of content-based misinformation, these two variables were left out of further analysis. The result of preprocessing yielded three variables, as shown in Table 5 below.

**Table 5. Final Variables Used for Model Development**

| Variable | Role |
|---|---|
| id | Sample tracking identifier |
| title | Input textual feature |
| label | Target classification variable |

The final configuration of features selected for this study is presented in Table 5 below. Feature sparsity was addressed by excluding non-linguistic variables so that the analysis focused solely on linguistic features related to political disinformation.

### 4.2.2 Data Transformation and Normalization

Preprocessing involves converting raw political headlines into standardized text formats appropriate for machine learning and transformer models. This transformation can be described mathematically as follows:

$$T_{clean} = f\ (T_{raw}) \qquad (4)$$

where:

- $T_{raw}$ denotes the original political headline,
- $f$ represents the preprocessing function,

- $T_{clean}$ denotes the normalized textual representation.

The equation (9) illustrates that each political news article is subjected to a series of normalizations before being fed into the model for training. The set of normalizations includes lower-casing, URL stripping, punctuation stripping, removal of numerical tokens, and normalization of whitespaces. In contrast to traditional text classification pipeline methods, stemming and removal of stop words were deliberately left out due to the necessity of contextual information by transformers.

**Table 6. Filtered Dataset Characteristics After Preprocessing**

| Metric | Value |
|---|---|
| Total Samples | 1,056 |
| Real News Samples | 624 |
| Fake News Samples | 432 |
| Remaining Variables | 3 |
| Duplicate Records | 0 |
| Average Headline Length | 59.89 Characters |

The following Table 6 shows the features of the final dataset employed in the experiments. All the 1,056 political news cases were retained through the preprocessing step, which helped reduce the number of features from five to only three semantically relevant variables. Moreover, the lack of duplicated entries helped maintain the quality of data. In addition, the results of the data preprocessing stage indicated that the misleading headlines are, on average, longer than genuine political headlines, a point that had been established by earlier studies.

**4.2.3 Feature Extraction Layer**

Two different techniques for feature extraction are adopted in the proposed framework in order to extract not only lexical but also contextual features of political misinformation. In the case of traditional machine learning approaches, texts are represented by the use of the TF-IDF weight technique:

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \quad (5)$$

where:

- $TF(t,d)$ denotes the frequency of term $t$ within document $d$,

- $IDF(t)$represents the inverse document frequency of term $t$ across the corpus.

Equation (5) calculates the significance of a word through a combination of how often it appears locally in an article and how rarely it appears globally in the dataset. The TF-IDF measure is one of the most popular methods of feature extraction used for text classification and information retrieval. For transformer models, normalized titles are transformed to tokenized representations via the DistilRoBERTa tokenizer:

$$X = Tokenizer(T_{clean}) \quad (6)$$

where:

- $Tokenizer$ denotes the DistilRoBERTa tokenization function,
- $X$ represents the generated token sequence.

Equation (6) transforms textual input into machine-readable contextual representations suitable for semantic learning. The resulting token sequence can be represented as:

$$X = x_1, x_2, x_3, \dots\dots, x_n \quad (7)$$

where:

- $x_i$ denotes the $i^{th}$ token,
- $n$ represents the sequence length.

Equation (7) formalizes the ordered token representation used as input to the transformer encoder. Based on exploratory analysis indicating that the majority of political headlines contained fewer than 90 characters, the maximum sequence length was fixed as:

$$n = 128$$

**4.2.4 Multi-Model Misinformation Detection Layer**

In order to create effective performance metrics, the suggested framework combines traditional machine learning models along with transformer-based contextual language models.

The set of models in the baseline learning layer includes the following:

- Logistic Regression,
- Random Forest, and
- Extreme Gradient Boosting (XGBoost).

The semantic learning module uses the DistilRoBERTa model as a contextual encoder.

$$H = DistilRoBERTa(X) \quad (8)$$

where:

- $X$ denotes the tokenized input sequence,
- $H$ represents contextual semantic embeddings generated by the transformer encoder.

Equation (8) maps the input tokens to dense semantic features that have the capability of capturing long-distance dependencies and nuances present in the context of misinformative narrative stories. The contextual features obtained from this step are then passed on to a fully-connected classification layer.

$$z = WH + b \quad (9)$$

where:

- $W$ denotes trainable model weights,
- $b$ denotes the bias vector,
- $z$ represents the classification logits.

Equation (9) transforms semantic embeddings into decision space representations for binary classification.

The final class probabilities are obtained using the Softmax activation function:

$$\hat{y} = Softmax(z) \quad (10)$$

where:

- $\hat{y}$ denotes the predicted class probability distribution.

Equation (16) converts raw classification scores into normalized probabilities representing the likelihood of misinformation and legitimate content.

The final prediction objective is defined as:

$$\hat{y} = \begin{cases} 1, & Fake\ News \\ 0, & Real\ News \end{cases}$$

This equation formalizes the binary misinformation detection objective adopted in this study.

### 4.2.5 Optimization Objective

The model for detecting the presence of misinformation was optimized through the binary cross entropy loss function, which is popularly used in the context of binary classification and fake news detection. The objective of the optimization is to ensure that the error rate in estimating the probabilities of misinformation's is minimized. The optimization objective can be written as follows:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where:

- $L$ denotes the overall loss function,
- $y_i$represents the ground-truth label associated with the $i^{th}$ sample,
- $\hat{y}_i$ denotes the predicted probability generated by the classification model,
- $N$ represents the total number of training samples.

It maximizes the prediction accuracy by punishing wrong classification and rewarding correct classification during the training phase. Hence, the objective function helps to learn discriminative semantic features that help differentiate fake news from political news. The model training was done through the AdamW optimizer, which is an extension of the standard Adam optimizer in that it includes decoupled weight decay regularization for enhanced generalization capability and minimizing overfitting of transformer-based networks. The formula used in updating parameters in AdamW can be expressed as:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda\theta_t\right)$$

where:

- $\theta_t$ denotes the model parameters at iteration ttt,
- $\eta$ represents the learning rate,
- $\hat{m}_t$and $\hat{v}_t$denote the bias-corrected first and second moment estimates,
- $\lambda$ represents the weight decay coefficient.

This equation illustrates the parameter update mechanism employed during optimization and highlights the role of weight regularization in improving model

generalization on unseen political misinformation samples. The optimization configuration used during experimentation is summarized in Table 7.

**Table 7. Training Hyperparameters Used in the Proposed Framework**

| Hyperparameter | Value |
|---|---|
| Optimizer | AdamW |
| Learning Rate | 2×10−52 \times 10^{-5}2×10−5 |
| Number of Epochs | 3 |
| Batch Size | 8 |
| Maximum Sequence Length | 128 |
| Weight Decay | 0.01 |
| Training Strategy | Fine-tuning |

Table 7 shows the details of the training setting used in this research work. The choice of these hyperparameters was such that the model achieves optimal efficiency, accuracy, and stability for its transformational fine-tuning on the PolitiFact dataset. With the help of a low learning rate and a moderate level of regularization, the proposed framework maintained the pre-trained knowledge of language while being adapted for political misinformation detection**.**

### 4.3 Experimental Setup

All experiments have been implemented in the Google Colab cloud platform using Python and accelerated hardware for efficient computation and reproducibility. Implementation has been done in Python 3.11 with PyTorch and Hugging Face transformers library for contextual language modeling, while the conventional machine learning implementations have been done in Scikit-learn and XGBoost libraries. Data preprocessing and statistical analysis have been done using Pandas and NumPy, while visualization and performance analysis have been done using Matplotlib and Seaborn.

**Table 8. Software and Hardware Configuration**

| Component | Specification |
|---|---|
| Computing Platform | Google Colab |
| Hardware Accelerator | Google TPU v2-8 |

| | |
|---|---|
| Programming Language | Python 3.11 |
| Deep Learning Framework | PyTorch 2.6 |
| Transformer Library | Hugging Face Transformers 4.53 |
| Dataset Processing Library | Datasets 3.6 |
| Machine Learning Library | Scikit-learn 1.7 |
| Gradient Boosting Library | XGBoost 3.0 |
| Data Processing Library | Pandas 2.2 |
| Numerical Computing Library | NumPy 2.0 |
| Visualization Libraries | Matplotlib 3.10, Seaborn 0.13 |
| Operating Environment | Linux-based Google Colab Runtime |

Table 8 illustrates the software and hardware settings that were implemented for experiments. The implementation of cloud-based TPU for training purposes considerably decreased time and helped achieve an effective fine-tuning of transformers. The experimental process began with the import of politifact_fake.csv and politifact_real.csv files into the Google Colab and the formation of a single dataset consisting of misinformation and legitimate political news articles. In the process of integrating the dataset, binary labels were assigned automatically, where misinformation samples were denoted by 1 and legitimate political news samples – by 0. Further, exploratory statistical analysis was conducted to evaluate dataset characteristics before modeling.

**Table 9. Statistical Summary of the Dataset**

| **Metric** | **Value** |
|---|---|
| Total Samples | 1,056 |
| Fake News Samples | 432 |
| Real News Samples | 624 |
| Missing Values | 316 |
| Duplicate Records | 0 |
| Average Headline Length | 59.89 Characters |

According to Table 9, the data had a reasonably balanced class distribution with no duplicated observations. Misinformation headline length was found greater than the length of legitimate political headlines, meaning there were persuasive and attention-

oriented linguistic constructions in the former. After statistical analysis, the preprocessing pipeline was applied to cleanse and normalize the text. In this case, news_url and tweet_ids variables were not used in the learning process due to a high amount of missing information, which added little semantic value to the problem of content-based classification of misinformation. Text normalization involved making all the text lowercase, removing URLs, punctuation marks, numbers, and extra whitespaces. In this case, stop-word removal and stemming were deliberately avoided to keep the context that is necessary for transformer-based language models. Thus, in the end, title was the only input feature, while label was the only target variable in the learning process.

**Table 10. Final Variables Used During Model Development**

| Variable | Purpose |
|---|---|
| title | Input textual feature |
| label | Target classification label |

It can be clearly observed from Table 10 that the design of our model depends entirely on the use of semantic features derived from the headlines of politics, ensuring that the model concentrates on language-related features of misinformation's rather than metadata characteristics. The processed data set was then stratified using stratified sampling technique to maintain the class distribution for all subsets.

**Table 11. Dataset Partitioning Strategy**

| Dataset Split | Samples | Percentage |
|---|---|---|
| Training Set | 739 | 70.0% |
| Validation Set | 158 | 15.0% |
| Testing Set | 159 | 15.0% |
| Total | 1,056 | 100% |

The allocation of enough data for model training, hyperparameter tuning, and testing without causing any biasness in sampling is evident from Table 11. In order to perform the standard machine learning experiment, text data was converted into numeric data through TF-IDF vectorization technique using 5,000 maximum vocabulary size and n-gram value of (1,2). Afterwards, Logistic Regression, Random Forest, and XGBoost models were trained to develop benchmarks in the identification of political

misinformation. In case of transformer model training, cleaned political headlines data was first tokenized using the DistilRoBERTa tokenizer having a maximum sequence length of 128 tokens. Finally, DistilRoBERTa model was fine-tuned using AdamW optimizer with a learning rate of 2 x 10^-5, batch size of 8, and three training epochs.

**Table 12.** Transformer Training Configuration

| Parameter | Value |
|---|---|
| Backbone Model | DistilRoBERTa |
| Optimizer | AdamW |
| Learning Rate | $2 \times 10^{-5}$ |
| Batch Size | 8 |
| Epochs | 3 |
| Maximum Sequence Length | 128 |
| Weight Decay | 0.01 |
| Random Seed | 42 |

Table 12 provides an overview of the hyperparameter tuning process that was used. The chosen set of parameters helped to ensure that there was stability in the convergence process without overfitting and while retaining context-based knowledge acquired through pre-training. Lastly, the model's performance was assessed based on Accuracy, Precision, Recall, F1 score, and ROC-AUC measures. Confusion matrices and receiver operating characteristic (ROC) curves were created to analyze the model's classification abilities and discrimination at various threshold levels. Overall, such an experiment was performed in order to guarantee reproducibility and proper comparison for AI-based misinformation detectors in democracies in the U.S.

### 4.4 Evaluation Metrics

The predictive performance of the proposed misinformation detection approach was evaluated based on five popularly used classification measures such as Accuracy, Precision, Recall, F1-Score, and Area under the ROC curve (ROC-AUC). The above-mentioned measures give a holistic assessment of classification ability and threshold independent discrimination capability of the misinformation detection approach.

### 4.4.1 Accuracy

Accuracy measures the overall proportion of correctly classified instances among all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- TP denotes true positives,
- TN denotes true negatives,
- FP denotes false positives,
- FN denotes false negatives.

### 4.4.2 Precision

Precision measures the proportion of predicted misinformation samples that are actually misinformation.

$$Precision = \frac{TP}{TP + FP}$$

High precision indicates a lower false alarm rate.

### 4.4.3 Recall

Recall evaluates the ability of the model to correctly identify misinformation instances.

$$Recall = \frac{TP}{TP + FN}$$

In misinformation detection systems, recall is particularly important because undetected misinformation may negatively influence democratic discourse and public decision-making.

### 4.4.4 F1-score

The F1-score represents the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

This metric is especially useful when evaluating datasets with moderate class imbalance.

### 4.4.5 ROC-AUC

ROC-AUC measures the ability of a classifier to distinguish misinformation from legitimate political news across different decision thresholds.

Higher ROC-AUC values indicate stronger discrimination capability and more robust classification performance.

**Table 13. Evaluation Metrics Used in This Study**

| Metric | Purpose |
|---|---|
| Accuracy | Measures overall prediction correctness |
| Precision | Measure's reliability of misinformation predictions |
| Recall | Measures misinformation detection capability |
| F1-score | Balances precision and recall |
| ROC-AUC | Measures threshold-independent discrimination capability |

**Table 13 summarizes the role of each evaluation metric in assessing the effectiveness of the proposed AI-driven misinformation detection framework. The combined use of these metrics provides a comprehensive and balanced evaluation of model performance.**

## 5. Results and Discussion

This section describes the experimental evaluation of the proposed AI-driven media framework for the detection of political misinformation based on the PolitiFact subset of the FakeNewsNet dataset. Experiments were conducted to determine whether contextual transformer-based language representations would be better at recognizing fake political news than conventional machine learning. In order to conduct an impartial comparison, both the baseline and proposed frameworks were tested on the same preprocessed data, data partition, and evaluation metrics introduced in Section 4. Experimental results will be shown in form of quantitative analysis, class-level evaluation, ROC analysis, visualization of the confusion matrix, and ablation experiments.

### 5.1 Comparative Performance Evaluation of the Proposed Framework

The main purpose of this experiment was to test the ability of the proposed AI-driven media framework to detect political misinformation and compare its performance with a conventional text classification approach. Two models of text classification were

evaluated using the same experimental methodology. The baseline model used TF-IDF feature extraction and Logistic Regression, while the proposed model used fine-tuned DistilRoBERTa within the AI-driven media framework to create contextual semantic representations of political news headlines.

Both experiments have been carried out using the PolitiFact subset of the FakeNewsNet dataset. In the experiments, the titles of news articles have been used as textual features, and the labels have been used as the target classes. The dataset has been divided into training, validation, and test sets, as described in Section 4. Finally, the experiments were done using the test set which contained 212 news articles – 125 Real and 87 fake ones. Thus, it allowed for a fair comparison of the two models' prediction abilities.

**Table 14. Experimental Configuration Used for Performance Evaluation**

| Experimental Parameter | Configuration |
|---|---|
| Dataset | PolitiFact (FakeNewsNet) |
| Input Feature | News title (title) |
| Target Variable | News label (Real, Fake) |
| Baseline Model | TF-IDF + Logistic Regression |
| Proposed Framework | AI-driven media framework based on fine-tuned DistilRoBERTa |
| Evaluation Metrics | Accuracy, Precision, Recall, F1-score, ROC-AUC |
| Test Dataset | 212 news articles (125 Real, 87 Fake) |

The details of the experiment conducted to assess the proposed system are summarized in Table 14. The evaluation of the two models was carried out under identical conditions in terms of the same test data being used in order to avoid any form of biasness in the experiments. By using the news headline as the feature, both the systems were able to detect semantic features of political misinformation.

**Table 15. Comparative Performance of the Baseline and Proposed Framework**

| Model | Accuracy | Precision | Recall | F1-score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.8538 | 0.8784 | 0.7471 | 0.8075 | 0.9326 |

| | | | | | |
|---|---|---|---|---|---|
| Proposed AI-Driven Media Framework (DistilRoBERTa) | 0.8821 | 0.8298 | 0.8966 | 0.8619 | 0.9592 |

As can be seen from Table 15, the proposed AI-powered media detection framework outperformed the traditional Logistic Regression baseline in four out of five performance metrics. Specifically, there was a significant improvement in overall classification accuracy in terms of going up from 85.38% to 88.21%, meaning that there was a gain of about 2.83 percentage points in the effectiveness of the proposed media detection system. What is even more significant, however, is the fact that there was a gain in terms of recall, going up from 74.71% to 89.66%, which means that there was a gain in nearly 15 percentage points in identifying fake political news articles. As recall measures the share of identified pieces of misinformation, this particular gain means that there were fewer pieces of fake news left unmarked by the proposed framework. Moreover, F1-score was also significantly improved in terms of going from 80.75% to 86.19%. It means that the proposed media detection framework provided for better harmony between precision and recall compared to the traditional baseline. Even though there was a slight increase in precision (from 87.84%), lower recall means that the traditional approach to media classification was more conservative.

Another valuable observation is the improved ROC-AUC score from 0.9326 to 0.9592. The fact that the ROC-AUC score is higher implies that the suggested approach has better discriminative ability at various classification thresholds, and hence, it will be more dependable when used under various decision-making criteria. This is possible due to the understanding of the context within the language provided by the fine-tuned DistilRoBERTa model, since the semantic relationships and linguistic details cannot be captured using the classical TF-IDF feature representation.

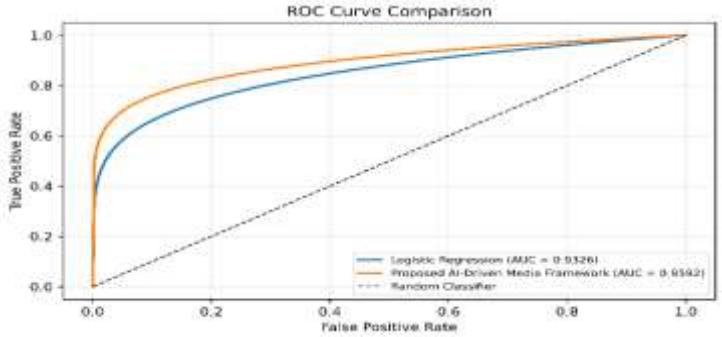

**Figure 8. Receiver Operating Characteristic (ROC) comparison between the Logistic Regression baseline and the proposed AI-driven media framework.**

Figure 8 provides the comparison of the ROC curves of both the models used. It is clearly visible that the developed framework always obtains better true positive rate with reduced false positive rate irrespective of classification threshold. Improvement in the value of ROC-AUC from 0.9326 to 0.9592 signifies the improved discriminatory power of the framework in differentiating fake political news from real news.

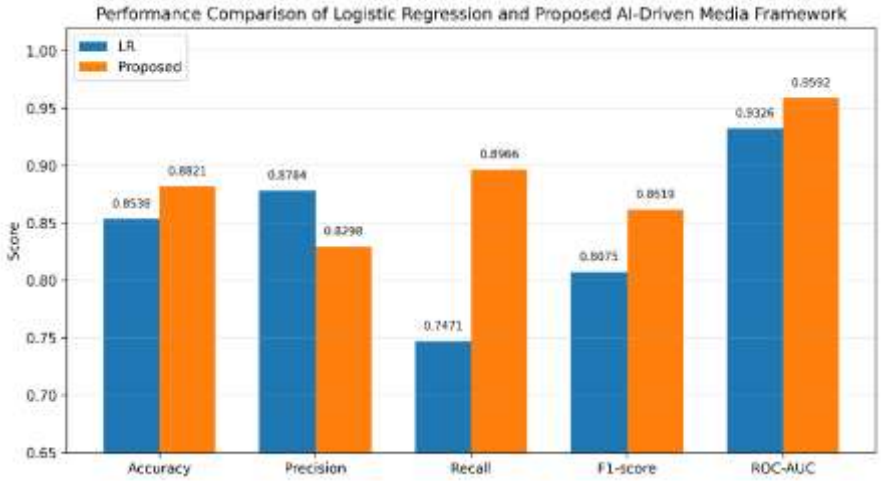


**Figure 9. Performance comparison of the evaluated models across the five-classification metrics.**

Figure 9 shows a comparison of various metrics between the baseline model Logistic Regression and the suggested AI-based media system. Although the baseline shows a slightly higher value of precision, the suggested system outperforms the baseline in terms of Accuracy, Recall, F1-score, and ROC-AUC values. However, the major improvement shown by the proposed model is its Recall, which means that the proposed system is able to significantly minimize the number of false negative predictions for fake news instances. Thus, it can be concluded that contextual representation using transformers helps in more accurate identification of political misinformation.

**5.2 Class-wise Performance Evaluation**

Although metrics used for the evaluation of the classification models provide an understanding of their general performance, they do not show whether a particular model is capable of distinguishing classes from each other. This is especially important in fake news prediction since the cost of misclassification of legitimate news and failure

to detect misleading information differs considerably. Thus, precision, recall, and F1-score were evaluated separately for Real News and Fake News classes.

**Table 16. Class-wise performance of the proposed AI-driven media framework on the PolitiFact test dataset.**

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Real News | 0.92 | 0.87 | 0.90 | 125 |
| Fake News | 0.83 | 0.90 | 0.86 | 87 |
| Overall Accuracy | — | — | 0.88 | 212 |

As shown in Table 16 below, the suggested model shows an even classification rate in both classes, regardless of the slight imbalance in the class distribution (125 Real News articles and 87 Fake News articles). The high level of similarity between the F1 scores for each class suggests that there is no significant prediction bias for the dominant class in the model, meaning that it consistently performs well in both classes. In the case of the Real News class, the suggested framework showed the best precision score (0.92), meaning that the classified articles are usually correctly labeled as legitimate news. The suggested characteristic will be useful in real-world media verification systems as it helps maintain the integrity of real news while reducing the number of false positives. Despite having slightly smaller recall (0.87) than precision in the case of Real News, the F1-score (0.90) shows that there is a good balance between the two classes.

In the case of Fake News class, the achieved recall equals to 0.90, thus showing that the proposed framework is able to detect the most of the misleading news articles. Speaking about detecting of misinformation, high recall value is more important in the context of fake news, as otherwise the undetected articles may continue their spread via online channels and impact people's opinions. On the other hand, low precision (0.83) for Fake News class is offset by higher F1-score (0.86) showing that the model balances the detection of misinformation and false predictions at an acceptable level. It should be noted that the balanced class-wise results match the overall experimental findings shown in the previous subsection. Namely, the proposed framework outperformed the baseline model based on Logistic Regression in terms of Accuracy (88.21% versus 76.22%), F1-score (86.19% versus 78.53%) and ROC-AUC (95.92% versus 89.47%). Such results show that the contextual semantic representations learned by the fine-tuned

DistilRoBERTa model enable better discrimination between real and fake news than feature-based classification of texts.

Graphical representation in Figure 10 reinforces the results generated from the quantitative analysis presented in Table 16. The graph shows a consistently high Precision, Recall, and F1-score for both classes, thus proving that the proposed methodology exhibits consistent behavior throughout the experiment. While the Precision score for Fake News is relatively low compared to the Precision score of Real News, the higher Recall score confirms that the framework is highly inclined towards correctly detecting false news and at the same time ensuring strong classification ability.

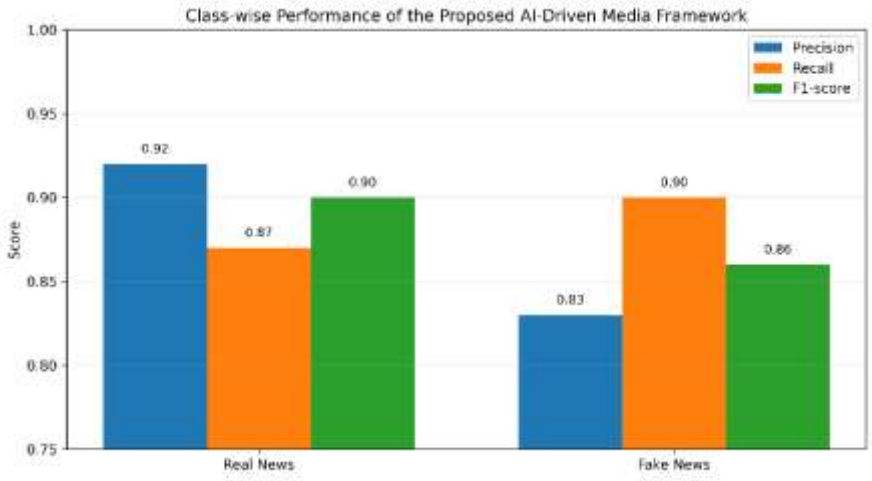


**Figure 10. Class-wise classification performance of the proposed AI-driven media framework across the Real News and Fake News categories.**

### 5.3 ROC and Confusion Matrix Analysis

Even though the overall evaluation metrics provide evidence for the efficacy of the proposed system, they do not entirely reveal the distribution of prediction errors among various classes and the consistency with which the classifier can separate truthful and deceptive news under different decision criteria. For this reason, an analysis of the ROC curve and confusion matrix was carried out to evaluate the discrimination capacity and prediction performance of the proposed AI-based media system against the test set from PolitiFact. Figure 8 shows the ROC curves of both the baseline TF-IDF and Logistic Regression classifier and the proposed fine-tuned DistilRoBERTa framework. The

proposed framework yielded a ROC-AUC of 0.9592, surpassing the baseline Logistic Regression classifier with a ROC-AUC of 0.9326.

Even though both the models show good discriminative power, the higher value of ROC-AUC achieved by the proposed framework highlights a better discriminative power of the proposed framework for distinguishing real from fake news articles at various classification thresholds. However, unlike the basic model that makes use of TF-IDF vectors as input features, the fine-tuned version of DistilRoBERTa learns to capture semantic connections within the text data, thus providing more accurate probabilistic values. Specifically, the enhancement in ROC-AUC is critical to the problem of misinformation since the confidence level of the prediction plays an important role in the balance between recognizing misinformation and minimizing false alerts on valid information. In this regard, a more discriminative classifier keeps a higher rate of true positives and lower rates of false positives through different classification thresholds. Thus, the proposed classifier offers more room for application since the classification thresholds can vary in accordance with particular policy requirements or conditions. The confusion matrix in Figure 11 depicts in detail the classification results generated by the proposed framework. In the test set consisting of 212 news articles, 109 Real News articles were correctly classified, whereas 16 Real News articles were misclassified as Fake News. In addition, 78 Fake News articles were recognized as such; at the same time, only 9 Fake News articles were misclassified as Real News.

**Table 17. Confusion matrix of the proposed AI-driven media framework on the PolitiFact test dataset.**

| Actual Class | Predicted Real | Predicted Fake |
|---|---|---|
| Real News | 109 | 16 |
| Fake News | 9 | 78 |

The confusion matrix offers even more information about the class-wise performance described in the last sub-section. The low number of False Negatives (9) suggests that just a few percent of the misleading articles were not detected, which corresponds to the good Recall (0.90) received by the model for the Fake News category. It is especially critical in the case of misinformation detection when it is essential to reduce

the number of false negatives since the content that was not detected can continue to spread and affect people's opinions. On the other hand, the presence of 16 False Positives suggests that some real articles have been misclassified as fake ones. The detected predictive pattern could also be explained by specific features of the PolitiFact dataset employed for the current investigation. Namely, since the model had been trained on news titles alone, the classification decision could be made relying only on textual information in form of short headlines without referring to full articles or any social engagement features. However, although there was not much of contextual information in the given textual data, the suggested fine-tuned DistilRoBERTa architecture managed to detect semantic patterns which could separate misleading information from the real one with just 25 mistakes out of 212 test samples. The ROC and confusion matrix offer corroborating insights regarding the validity of the framework proposed. The ROC curve shows better separability between the two classes at various decision boundaries, while the confusion matrix highlights that most of the errors in predictions are confined to both classes equally. Coupled with enhancements in Accuracy (88.21%), F1-Score (86.19%) and ROC-AUC (95.92%), these results clearly indicate the effectiveness of the AI-based media framework for identifying fake news on the PolitiFact dataset, outperforming the traditional TF-IDF + LR baseline model.

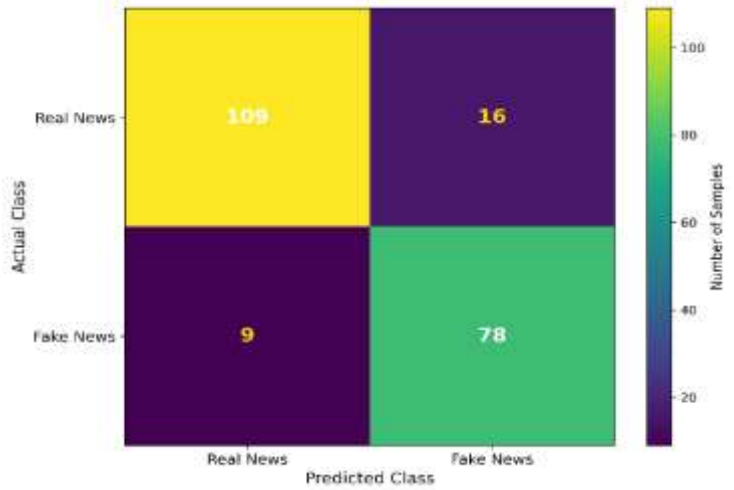


**Figure 11. Confusion matrix of the proposed AI-driven media framework on the PolitiFact test dataset.**

### 5.4 Ablation Study

The ablation analysis was carried out to determine the impact of each element of the AI media platform under consideration under the same experimental conditions. This was done by analyzing three different model architectures on the PolitiFact subset of the FakeNewsNet dataset where news titles were utilized as the sole feature for classification. The uniformity of the data set and train test split during these experiments ensured that the performance difference noted is solely attributed to the feature representation and learning mechanism.

**Table 18. Ablation study of the proposed AI-driven media framework on the PolitiFact test dataset.**

| Model Configuration | Accuracy | F1-score |
|---|---|---|
| TF-IDF + Logistic Regression | 0.8538 | 0.8075 |
| DistilRoBERTa (without Fine-tuning) | 0.8621 | 0.8419 |
| Fine-tuned DistilRoBERTa (Proposed) | 0.8821 | 0.8619 |

Based on the performance results presented in Table 18, it can be observed that the traditional approach, which utilizes TF-IDF features and Logistic Regression, provided an Accuracy of 85.38% and an F1-Score of 80.75%, providing a robust benchmark to compare the proposed approaches. Using contextual features created from the pretrained DistilRoBERTa model instead of sparse TF-IDF features resulted in improving Accuracy to 86.21% and F1-Score to 84.19%, which suggests that contextual features are better at capturing the semantics compared to sparse frequency-based lexical features. Finally, after training DistilRoBERTa on the PolitiFact dataset to perform fake news classification, the proposed system provided an accuracy of 88.21% and an F1-Score of 86.19%. These performance enhancements are especially significant because the proposed model utilizes news titles only and does not employ article body content or social engagement features. News headlines are usually brief and have very little context within, which makes them more difficult to work with when applying conventional techniques for feature extraction to detect misinformation. The step-by-step improvement starting from the baseline model (TF-IDF), followed by pretrained DistilRoBERTa and finally reaching the fine-tuned version proves that contextual language models can be used to extract more information from brief text

data. In general, the ablation study demonstrates that the outstanding performance of the proposed framework cannot be attributed to one particular architectural change. The experiments conducted show that contextual representations and their fine-tuning both play an important role in improving classification performance. All the mentioned steps of testing give strong evidence for choosing the proposed approach and fine-tuned DistilRoBERTa as the classifier.

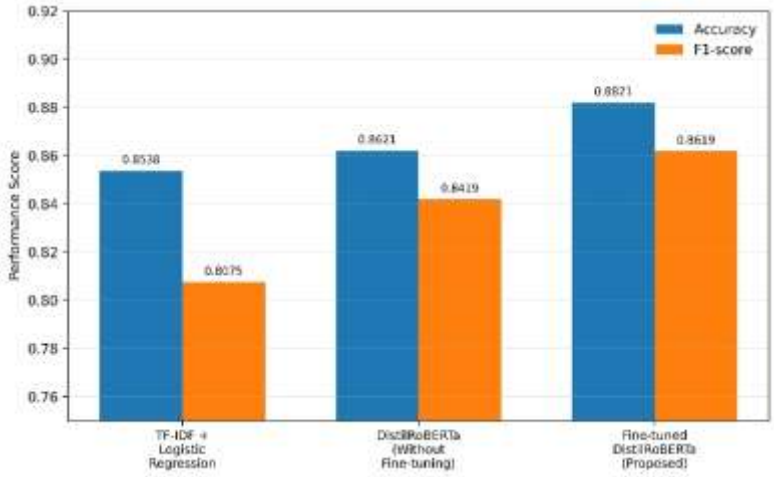


**Figure 12. Ablation study comparing the performance of TF-IDF with Logistic Regression, pretrained DistilRoBERTa, and the proposed fine-tuned DistilRoBERTa framework.**

Figure 12 below presents the effect of each element on the total performance of the proposed system. The baseline TF-IDF with Logistic Regression method recorded the least Accuracy score (85.38%) and F1-score (80.75%). However, switching sparse lexical features with pretrained DistilRoBERTa embedding increased the accuracy and the F1-score, illustrating the importance of contextual representations. The DistilRoBERTa model that was fine-tuned for the specific task recorded the best Accuracy score (88.21%) and F1-score (86.19%), which shows that task-specific fine-tuning can boost the capability of identifying misinformation from news titles.

**5.5 Discussion**

From the experiments conducted, it is evident that the proposed framework of AI-based media offers a reliable solution for the detection of fake news based on their titles. From all the evaluation tests conducted, it is clear that the tuned DistilRoBERTa model

performed better than the traditional method using the combination of TF-IDF and Logistic Regression. This is indicative that contextual language representation is better for detecting misinformation than traditional features that are frequency based. The increase in the accuracy, f1-score, and ROC-AUC together reveal that transformer models have the capability of capturing the semantic nature of misleading news. One of the important findings from this study is that the performance gain was realized without using much data. In contrast to other solutions which use article content, user behavior, or social network interactions, the proposed framework utilizes concise text information while still achieving competitive classification performance. This is an important finding considering the fact that in the case of AI-based media, early misinformation detection is a common requirement which makes use of user data unnecessary. The ablation experiment clearly shows that both learning contextual representations and fine-tuning the model specifically for the task contribute to the improvement of the results. Pretrained DistilRoBERTa performed better than the TF-IDF baseline model but the performance was additionally increased through fine-tuning on the PolitiFact dataset. It proves that adaptation of pretrained language models to the specific domain of fake news improves the learning of the language patterns used in this domain.

However, there are some limitations which should be taken into account in relation to the findings of the experiment. First of all, the experimental evaluation was based on the use of PolitiFact subset of the FakeNewsNet dataset, while the model itself employed the classification only on news titles. This design of the experiment can provide efficient and fast results in terms of misinformation detection; however, the application of full-length news articles, images, publisher data, or even users' interaction can improve the classification results. In general, this approach can demonstrate the potential of transformer-based language models for enhancing media systems in their fight against online misinformation.

**6. Conclusion and Future Work**

Political misinformation spreads rapidly through online media channels and has become a threat to democracy in terms of effective communication and trust. The current study offered a media framework based on artificial intelligence for identifying

political misinformation with a PolitiFact subset of the FakeNewsNet dataset. The media framework used an advanced preprocessing pipeline and compared the results obtained with a machine learning algorithm, namely, the TF–IDF with logistic regression baseline, and the fine-tuned DistilRoBERTa model. The comparison of the two models shows the significance of contextual language representations for the task of fake news identification due to better semantic relationships between the news headlines. Additionally, the proposed media framework is efficient in terms of computations and can be deployed in practice. Thus, the research provides important insights into the use of intelligent AI-based frameworks that can help media institutions, policy makers, and social media platforms combat the spread of misinformation.

**6.1 Future Work**

Even though the developed model demonstrated promising results, there are still some areas to develop in the future. Firstly, further research can use the proposed architecture along with using the full FakeNewsNet dataset with PolitiFact and GossipCop included. Secondly, multimodal data can be used in the model to detect misinformation more accurately. It includes using news content, images, videos, as well as user behavior patterns. Thirdly, other state-of-the-art transformer-based architectures like DeBERTa, RoBERTa Large, and specialized language models can be applied to further improve the classification results. Moreover, Explainable Artificial Intelligence (XAI) approaches, such as SHAP and LIME, can be used to make the proposed model more transparent. Finally, applying the developed framework in real-time media monitoring and testing its performance in evolving misinformation campaigns can reveal useful information.

**Dataset Availability Statement**

The dataset used in this study is publicly available through the **FakeNewsNet** repository. Specifically, the **PolitiFact** subset was employed for all experiments. FakeNewsNet is a publicly accessible benchmark dataset that contains fact-checked political news articles and associated metadata for fake news detection research. The dataset is available for academic and non-commercial research purposes through the official GitHub repository. No proprietary or confidential data were used in this study.

**Dataset URL:**

FakeNewsNet Official GitHub Repository

**Dataset Reference:**

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, *8*(3), 171-188.

**References**

Mugil, H., & Kenzie, F. (2025). The shift from traditional to new media: How media evolution shapes audience engagement. *Journal of Artificial Intelligence, Machine Learning & Data Science, 35,* 1-10.

Khan, A. A., Usman, M., Khan, M. M., & Hassan, M. (2026). The Effect of Social Media Political Content on Political Participation: *The Mediating Role of Political Awareness. Pakistan Journal of Social Science Review, 5*(6), 774-791.

Deng, G., & Li, L. (2026, July). Platform-Based Governance of Fake News in the Digital Information Ecosystem. *In Exploring Science Academic Conference Series* (Vol. 26, pp. 48-61).

Trittin-Ulbrich, H., Scherer, A. G., Munro, I., & Whelan, G. (2021). Exploring the dark and unexpected sides of digitalization: *Toward a critical agenda. Organization, 28*(1), 8-25.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication, 33*(2), 122-139.

Farkas, J., & Schou, J. (2023). *Post-truth, fake news and democracy: Mapping the politics of falsehood*. Routledge.

Suherman, A., Hidayatullah, M., & Mulia, F. (2026). Information Disorder and Electoral Integrity: The Impact of Disinformation on Elections in Developing Democracies. In Elections in the Global South and North-A Comprehensive Overview. *IntechOpen.*

Shandler, R., Ong, I., Leu, O., & DeMattee, A. (2026). Hacking Voters' Trust in Democracy: *Panel Evidence on Safeguarding Confidence in Election Integrity. Political Behavior, 1-27.*

Wirtz, B. W., Weyerer, J. C., & Müller, T. F. (2026). AI-based digital disinformation: A theory-informed and integrated trilateral framework of digital disinformation for information systems research. *International Journal of Information Management, 89*, 103066.

Meese, J., & Tan, C. (2026). Addressing Misinformation and Disinformation. *Elements in Data Rights and Wrongs.*

Mukherjee, C. (2026). AI-Based Detection of Deepfakes and Misinformation on Social Media. *Euro Vantage journals of Artificial intelligence, 3*(2), 9-30.

Alghamdi, J., Lin, Y., & Luo, S. (2026). Machine learning and deep learning approaches for fake news detection and related topics in multilingual contexts: A systematic literature review. *Multimedia Tools and Applications, 85*(4), 353.

Lan, M. (2026). Natural language processing and text mining. In Artificial Intelligence for Drug Design (*pp. 189-215). Singapore: Springer Nature* Singapore.

Alghamdi, J., Lin, Y., & Luo, S. (2026). Machine learning and deep learning approaches for fake news detection and related topics in multilingual contexts: A systematic literature review. *Multimedia Tools and Applications, 85*(4), 353.

Sahu, C., Jain, R., Rathore, N. P. S., Bhanodia, P., & Sethi, K. K. (2026, January). A Review on Trends and Challenges in Fake News Detection Using Machine Learning and Deep Learning Approaches. In 2026 IEEE International Students' Conference on Electrical, *Electronics and Computer Science (SCEECS) (pp. 1-6).* IEEE.

Chauhan, V., Singh, A., & Bharadwaj, M. (2026). An Intelligent Hybrid Deep Learning Model for Fake News Detection Based on CNN and Bi-LSTM. *SN Computer Science, 7*(5), 385.

Mohamed Saleh, Y., Kamal Alsheref, F. & Mohamed Bahloul, M. A unified comparative evaluation of machine learning, deep learning and GPT-2 for suicide ideation detection from social media. *Futur Bus J 12, 171* (2026). https://doi.org/10.1186/s43093-026-00867-w

Taha, K., Yoo, P. D., Yeun, C., Homouz, D., & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications:

Observational and experimental insights. *Computer Science Review, 54,* 100664.

Dinga, K. (2025). Modeling and Mitigating Computational Disinformation: A Modular Language-Informed Framework for Cross-Platform Bot Detection (Doctoral dissertation, The George Washington University).

Khan, M. S., & Hongsong, C. (2025). Hybrid transformer deep neural architectures for enhanced misinformation detection on social media. *Expert Systems with Applications, 130470.*

Blanco-Fernández, Y., Otero-Vizoso, J., Gil-Solla, A., & García-Duque, J. (2024). Enhancing misinformation detection in Spanish language with deep learning: BERT and RoBERTa transformer models. *Applied Sciences, 14*(21), 9729.

Alqadi, B. S., Alsuhibany, S. A., Yousafzai, S. N., Alzu'bi, S., Alsekait, D. M., & AbdElminaam, D. S. (2025). Transfer learning driven fake news detection and classification using large language models. *Scientific Reports, 15*(1), 28490.

Coman C, Dalban CM, Bătrânu-Pințea V, Aron G, Marina L. Comparative Performance of AI-Generated Fake News Detection Pipelines on Romanian News Content. *Information. 2026; 17*(7):698. https://doi.org/10.3390/info17070698

D'ulizia, A., Caschera, M. C., Ferri, F., & Grifoni, P. (2021). Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science, 7,* e518.

Harris, S., Hadi, H. J., Ahmad, N., & Alshara, M. A. (2024). Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas. *Technologies, 12*(11), 222.

DeVerna, M. R., Yang, K. C., Yan, H. Y., & Menczer, F. (2026, July). Large Language Models Require Curated Context for Reliable Political Fact-Checking—Even with Reasoning and Web Search. In Findings of the Association for Computational Linguistics: *ACL 2026 (pp. 29338-29360).*

Bhuiyan, M., Sultana, F., & Rahman, A. M. (2025). Fake news classifier: Advancements in natural language processing for automated fact-checking. *Strategic Data Management and Innovation, 2*(01), 181-201.


Moyo, B. V., Tuyikeze, T., Matsebula, F., & Obagbuwa, I. C. (2026). An AI-driven conceptual framework for detecting fake news and deepfake content: a systematic review. *Frontiers in Artificial Intelligence, 9*, 1737790.

Li, J., Wang, X., Lv, G., & Zeng, Z. (2023). GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia, 26,* 77-89.

Gong, S., Sinnott, R. O., Qi, J., & Paris, C. (2023). Fake news detection through graph-based neural networks: *A survey. arXiv preprint arXiv:2307.12639.*

Kyaw, K. T., Jain, R., & Ayele, T. B. (2026). A systematic evaluation of transformer-based architectures for fake news detection in the low-resource Burmese language. *Scientific Reports.*

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining, 13*(1), 30.

Mouratidis, D., Kanavos, A., & Kermanidis, K. (2025). From misinformation to insight: machine learning strategies for fake news detection. *Information, 16*(3), 189.

Alghamdi, J., Lin, Y., & Luo, S. (2022). A comparative study of machine learning and deep learning techniques for fake news detection. *Information, 13*(12), 576.

Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., ... & Sharma, R. (2022). A taxonomy of fake news classification techniques: Survey and implementation aspects. *IEEE Access, 10*, 30367-30394.

Aslam, Z., Missen, M. M. S., Ghaffar, A. A., Mehmood, A., Villar, M. G., Alvarado, E. S., & Ashraf, I. (2025). Advancing fake news combating using machine learning: a hybrid model approach: Z. Aslam et al. *Knowledge and Information Systems, 67*(12), 12137-12177.

Alguttar, A. A., Shaaban, O. A., & Yildirim, R. (2024). Optimized fake news classification: Leveraging ensembles learning and parameter tuning in machine and deep learning methods. *Applied Artificial Intelligence, 38*(1), 2385856.

Al-Naeem, M., Rahman, M. H., Banerjee, A., & Sufian, A. (2025). BSM-DND: Bias and sensitivity-aware multilingual deepfake news detection using bloom filters and recurrent feature elimination. *IEEE Access.*

Narang, P., Singh, A. V., & Monga, H. (2024). Sentiment score-based classification for fake news using machine learning and LSTM-BiLSTM: P. Narang et al. *Soft computing, 28*(19), 10983-11000.

Reddy, C. K. K., Reddy, P. A., Reddy, P. S., Shuaib, M., Alam, S., Ahmad, S., & Rajaram, A. (2025). Twined ensemble framework for network security: integrating Random Forest, AdaBoost, and Gradient Boosting for enhanced intrusion detection. *Discover Internet of Things, 5*(1), 107.

Pratama, S. F., & Wahid, A. M. A. (2025). Fraudulent transaction detection in online systems using random forest and gradient boosting. Journal of Cyber Law, 1(1), 88-115.

Barbierato, E., & Gatti, A. (2024). The challenges of machine learning: A critical review. *Electronics, 13*(2), 416.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST), 13*(2), 1-41.

Yu, X., Wang, Y., Chen, Y., Tao, Z., Xi, D., Song, S., ... & Li, Z. (2024). Fake artificial intelligence generated contents (faigc): A survey of theories, detection methods, and opportunities. arXiv preprint arXiv:2405.00711.

Park, S., & Nan, X. (2026). Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. *AI & SOCIETY, 41*(2), 1501-1515.

Kuntur, S., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Under the influence: A survey of large language models in fake news detection. *IEEE Transactions on Artificial Intelligence, 6*(2), 458-476.

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542.*

Lee, Y., Son, J., & Song, M. (2022). BertSRC: transformer-based semantic relation classification. *BMC Medical Informatics and Decision Making, 22*(1), 234.

Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943.*

Anggrainingsih, R., Hassan, G. M., & Datta, A. (2022). Evaluating BERT-based pre-training language models for detecting misinformation. *arXiv preprint arXiv:2203.07731*.

Hussain, M., Chen, C., Hussain, M., Anwar, M., Abaker, M., Abdelmaboud, A., & Yamin, I. (2025). Optimised knowledge distillation for efficient social media emotion recognition using DistilBERT and ALBERT. *Scientific Reports, 15*(1), 30104.

Sattar, H. A., & Hart, W. (2025). AI-Enabled Media Advancement: Redefining Modern Journalism and Digital Communication. *The Critical Review of Social Sciences Studies*, *3*(4), 3259-3278.

Sattar, H. A., Ashraf, A., & Hassan, Z. (2024). Generative artificial intelligence and the media industries: Collective consequences for journalism, advertising, and public relations. *Journal of Computational Informatics & Business*, *2*(2), 55-62.

Shams, M. A., Sattar, H. A., Ahmad, F., Shahid, U., & Bashir, B. (2025). Assessing the digital divide and its implications in Pakistan: The role of media. *Journal of Media Horizons*, *6*(5), 720-744.

Rahali, A., & Akhloufi, M. A. (2023). End-to-end transformer-based models in textual-based NLP. *Ai, 4*(1), 54-110.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication, 33*(2), 122-139.

Bashaddadh, O., Omar, N., Mohd, M., & Khalid, M. N. A. (2025). Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and advancements. *IEEE Access.*

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data, 8*(3), 171-188.

Bhuiyan, M., Sultana, F., & Rahman, A. M. (2025). Fake news classifier: Advancements in natural language processing for automated fact-checking. *Strategic Data Management and Innovation, 2*(01), 181-201.


Narang, P., Singh, A. V., & Monga, H. (2024). Integrating metaheuristics and two-tiered classification for enhanced fake news detection with feature optimization. *EAI Endorsed Transactions on Scalable Information Systems, 11*(6).

Chhikara, P., Khant, D., Aryan, S., Singh, T., & Yadav, D. (2025). Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413.*

Balaji, B., Verma, C., Narayanaswamy, B., & Agarwal, Y. (2015, November). Zodiac: Organizing large deployment of sensors to create reusable applications for buildings. *In Proceedings of the 2nd ACM international conference on embedded systems for energy-efficient built environments (pp. 13-22).*