A Comprehensive and Explainable Evaluation of Embedding Strategies and Large Language Models for COVID-19 Fake News Detection

Diya Prakash[1], Praveen Kumar S[1], Ranjith Kumar R[1], Siranjeevi Rajamanickam[2], and Balasubramanian Palani[1]

[1] Department of CSE, IIIT Kottayam, Kottayam, India
diyaprakash2205@gmail.com, spraveenkumar2205@gmail.com,
ranjith23bcd30@iiitkottayam.ac.in, pbala@iiitkottayam.ac.in
[2] Department of Computer Science, Government Polytechnic College,
Thuvakudimalai, Trichy, India
rajasiranjeevi@gmail.com

Abstract. Fake news dissemination poses serious risks in the healthcare domain, where misinformation can influence public behavior and decision making. Although recent NLP models have improved detection accuracy, many operate as black boxes, limiting trust and deployment. This paper presents an explainable fake news detection framework spanning lexical feature-based models, transformer architectures, and lightweight large language models. Classical machine learning methods, domain-specific transformers, and compact large language models are evaluated under a unified experimental setting. Explainable artificial intelligence techniques are applied at each stage to analyze decision behavior. Results show that large language models achieve the highest performance, while explainability analysis reveals a progression from lexical cues to contextual and reasoning-driven predictions.

Keywords: Fake News Detection · COVID-19 · Natural Language Processing · Transformers · Large Language Models · Explainable Artificial Intelligence

1 Introduction

The rapid dissemination of fake news has emerged as a critical challenge in the digital era, particularly in the healthcare domain, where misinformation can directly influence public behavior, policy decisions, and public trust. During large-scale health emergencies such as the COVID-19 pandemic, misleading content propagated through online news platforms and social media has demonstrated the potential to cause significant societal harm. Consequently, automated fake news detection has become an important research problem within natural language processing and machine learning.

Early approaches to fake news detection primarily relied on word frequency-based representations, including Bag-of-Words, TF-IDF, and HashingVectorizer, combined with traditional machine learning classifiers. While these methods are computationally efficient and relatively interpretable, their dependence on surface-level lexical patterns limits their ability to model semantic meaning and contextual dependencies, particularly for nuanced or well-crafted misinformation.

Transformer-based architectures addressed these limitations by generating contextualized representations that capture long-range dependencies and semantic relationships within text. Models such as BERT and domain-specific variants like MedBERT have demonstrated strong performance in health-related misinformation detection by leveraging contextual understanding and domain knowledge. More recently, large language models (LLMs) have introduced reasoning-driven approaches that evaluate global narrative coherence and semantic plausibility, enabling robust detection of subtle and context-dependent fake news.

Despite improved predictive performance, many fake news detection systems operate as black boxes, limiting trust in high-stakes domains such as healthcare. This motivates the integration of explainable artificial intelligence across representation paradigms. Accordingly, this paper presents an explainable fake news detection framework that evaluates frequency-based models, static embeddings, contextual transformers, and lightweight large language models under a unified setting.

2 Related Work

Detecting fake news has become a critical area of research in the digital age, with various methods emerging to tackle the spread of misinformation. These techniques can be broadly categorized into several key approaches, each with its own strengths and weaknesses. The evolution of these methods reflects the increasing sophistication of both fake news itself and the technologies used to identify it.

2.1 Word frequency encoding based FND methods

Early approaches to fake news detection often relied on word frequency encoding, a method that analyzes the statistical properties of text. These techniques operate on the principle that fake news articles may exhibit different linguistic patterns compared to genuine news. By counting the occurrences of specific words or phrases, these methods aim to identify signals of deception.

One common technique in this category is Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF evaluates how important a word is to a document in a collection of documents. This method helps in highlighting words that are frequent in a specific fake news article but not common across a larger corpus of genuine news, potentially indicating sensationalism or a narrow focus. For instance, a fake news article might overuse emotionally charged words or specific, misleading terminology. In their work, Cao et al. [1] utilized TF-IDF and CountVectorizer as part of their framework for extracting keyword information.

Similarly, Islam et al. [4] noted that their proposed classifier integrates features like TF-IDF and Bag of Words (BoW) for machine learning algorithms.

While these methods are computationally efficient and can be effective for certain types of fake news, they have significant limitations. They often fail to capture the nuances of language, such as context, sarcasm, and irony, which are crucial for understanding the veracity of a news story.

2.2 Pretrained static word embedding based FND methods

To address the lack of contextual understanding in word frequency methods, researchers turned to pretrained static word embeddings. These techniques represent words as dense vectors in a continuous vector space, where words with similar meanings are located closer to one another. Popular pretrained models include Word2Vec, GloVe, and FastText.

These embeddings are "static" because each word has a single vector representation, regardless of its context in different sentences. For example, the word "bank" would have the same vector in "river bank" and "investment bank." Despite this limitation, these models capture semantic relationships between words, which is a significant improvement over simple word frequency counts.

In fake news detection, these embeddings are used to convert the text of a news article into a series of vectors, which are then fed into neural networks like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) for classification. This approach allows the model to learn more complex patterns in the language used in fake and real news. For example, a study by Kishwar and Zafar, as cited by Islam et al. [4], developed a novel dataset for detecting fake news in Pakistan and found that an LSTM model with GloVe embeddings emerged as the top performer, achieving a notable F1-score of 0.94.

2.3 Contextual word embedding based FND methods

The limitation of static word embeddings led to the development of contextual word embedding models, which represent a major leap forward in natural language processing. These models, most notably BERT (Bidirectional Encoder Representations from Transformers) and its variants like RoBERTa, DeBERTa, and XLNet, generate different embeddings for a word depending on its context within a sentence. This is achieved through the use of the Transformer architecture, which employs self-attention mechanisms to weigh the importance of different words in a sentence when encoding a specific word.

In the context of fake news detection, contextual embeddings have proven to be highly effective. By fine-tuning these pretrained models on fake news datasets, researchers have been able to achieve state-of-the-art performance. These models are better equipped to understand the subtle linguistic cues, inconsistencies, and manipulative language often found in fake news.

Several recent papers highlight the power and challenges of using these models:

Chen et al. in "Real-time Factuality Assessment from Adversarial Feedback" [2] evaluated LLM-based detectors like GPT-4o and Gemini Pro on real-time news from sources like PolitiFact and Snopes, finding that while they perform well, their accuracy can be challenged by adversarially generated content. Their work introduces a method to create deceptive news variants that decrease a strong GPT-4o detector's ROC-AUC by an absolute 17.5%.

Mukherjee and Ghosh in "UNITE-FND" [6] propose a novel framework that reframes multimodal fake news detection as a unimodal text classification task. They use Gemini 1.5 Pro to convert visual content into structured text, which is then classified by models like BERT and DeBERTa, achieving 92.52% accuracy while reducing computational costs by over 10x compared to state-of-the-art multimodal models.

Cao et al. [1] proposed the SLIM framework, which detects fake news using systematically selected, limited information instead of full articles. Using XLNet as the encoder, their framework achieved performance comparable to state-of-the-art methods on datasets like ReCOVery, with an accuracy of 95.55%.

Islam et al. [4] developed the first large benchmark FND dataset for the Urdu language and proposed a unified LLM model that outperforms standalone models like XLNet, mBERT, and RoBERTa. Their stacked model, combining mBERT, XLNet, and XLM-ROBERTa, achieved an impressive accuracy of 0.956.

Wu et al. [**?**] introduced SheepDog, a style-robust fake news detector designed to defend against LLM-empowered style attacks. They found that such attacks could decrease the F1 score of state-of-the-art detectors by up to 38%. SheepDog uses a RoBERTa backbone and leverages LLMs for style-agnostic training to prioritize content over style.

Hu et al. in "Bad Actor, Good Advisor" [3] investigated the role of LLMs and found that while a sophisticated model like GPT-3.5 can provide valuable multi-perspective rationales for why a story is fake, it underperforms a fine-tuned SLM like BERT in direct classification. This led them to propose the ARG network, where an SLM is guided by rationales from an LLM, outperforming LLM-only and SLM-only baselines.

Xu and Li [10] conducted a comparative study showing that online LLMs like GPT-4, Claude, and Gemini are better suited for detecting emerging fake news in real-time than traditional offline models, which struggle with the dynamic nature of misinformation. Their experiments showed that an offline RoBERTa-based model (MDFEND) achieved a 0.557 F1 score on real-time news, while online models like Llama 3.1 achieved a 0.914 F1 score.

Su et al. [8] studied how to adapt fake news detectors to the era of LLMs by evaluating models on a mix of human-written and machine-generated news. A key finding was that detectors trained exclusively on human-written articles (using models like RoBERTa) could effectively detect machine-generated fake news, but the reverse was not true.

Yan et al. [9] proposed the BTCM model, which uses a 1D-CCNet attention mechanism for multimodal fake news detection. For the text component, the model utilizes a BERT encoder to extract features from news text, achieving an accuracy of 0.925 on the Twitter dataset.

Li et al. in their study "Fake news detection through topic and sentiment analysis based on fusion of graph convolutional network and BERT" [5] propose a method that combines graph convolutional networks (GCN) with BERT embeddings. They found that BERT effectively captures deep semantic information, and their fusion model outperformed standalone BERT on datasets, achieving an accuracy of 96.6% on the PolitiFact dataset.

2.4 Large Language Models and Explainable Artificial Intelligence for Fake News Detection

Recent studies have explored the application of large language models for fake news detection in dynamic and real-world settings. Chen *et al.* evaluated large language model based detectors such as GPT-4o and Gemini Pro on real-time fact-checking datasets and analyzed their robustness under adversarial conditions. Xu and Li compared online large language models, including GPT-4, Claude, Gemini, and LLaMA variants, with offline transformer-based models and demonstrated superior performance of online models in detecting emerging fake news.

Hybrid approaches combining large language models with smaller task-specific models have also been investigated. Hu *et al.* studied the use of large language models as rationale generators to guide smaller language models for fake news detection, showing improved performance over standalone models. Mukherjee and Ghosh proposed a framework that converts multimodal content into structured text using a large language model and performs classification using transformer-based models, achieving high accuracy with reduced computational cost.

Explainable Artificial Intelligence has been increasingly incorporated into fake news detection systems to analyze model behavior and decision-making. Prior works have applied post-hoc explanation techniques such as SHAP and LIME to traditional and deep learning models to interpret feature and token contributions. Recent studies in *Engineering Applications of Artificial Intelligence* and Springer journals have focused on explainability for deep and transformer-based models, highlighting the limitations of traditional attribution methods for large language models and motivating the use of rationale generation and perturbation-based explanation strategies.

3 Research Objectives

The primary objective of this study is to systematically analyze fake news detection approaches across different representation learning paradigms and model complexities. The work aims to evaluate the effectiveness of word frequency based methods, static word embeddings, contextual transformer models, and lightweight large language models under a unified experimental setting.

Another objective is to examine how model performance evolves as representation learning shifts from surface level lexical features to contextual and reasoning driven language models. By comparing traditional machine learning models with transformer based and large language models, the study seeks to understand their relative strengths in capturing semantic and contextual information for fake news detection.

In addition, this research aims to integrate explainable artificial intelligence techniques into the fake news detection pipeline to analyze decision making behavior across different model families. The objective is to assess model transparency by identifying influential features, tokens, and reasoning patterns that contribute to predictions.

Finally, the study seeks to investigate the trade offs between classification accuracy and interpretability, particularly in the context of large language models, and to highlight the importance of explainability for building trustworthy and reliable fake news detection systems in the healthcare domain.

4 Proposed Methodology

This section describes the proposed explainable fake news detection framework, which enables a systematic comparison of different representation learning paradigms under a unified experimental setting. The framework follows a multi-stage architecture comprising preprocessing, representation learning, classification, and explainability analysis.

4.1 Overview

The complete workflow of the proposed framework is illustrated in Figure 1. The process begins with raw textual data collected from COVID-19–related news articles and social media posts. The text is first subjected to standard preprocessing operations, including lowercasing, stopword removal, tokenization, stemming, and lemmatization, to reduce noise and ensure linguistic consistency across the dataset.

The framework follows a multi-stage architecture that progressively evaluates fake news detection models with increasing representational capacity. In the first stage, the preprocessed text is encoded using word frequency based representations such as Bag-of-Words, Term Frequency–Inverse Document Frequency, and HashingVectorizer. These representations are supplied to classical machine learning classifiers, including Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting, to establish interpretable baseline performance.

In the second stage, the framework incorporates pretrained static word embeddings to capture semantic relationships beyond surface-level lexical patterns. Word2Vec, GloVe, and FastText embeddings are used to transform textual data into dense vector representations, which are then provided as input to neural network based classifiers for fake news detection. The third stage employs contextual transformer based models to generate context-aware representations of

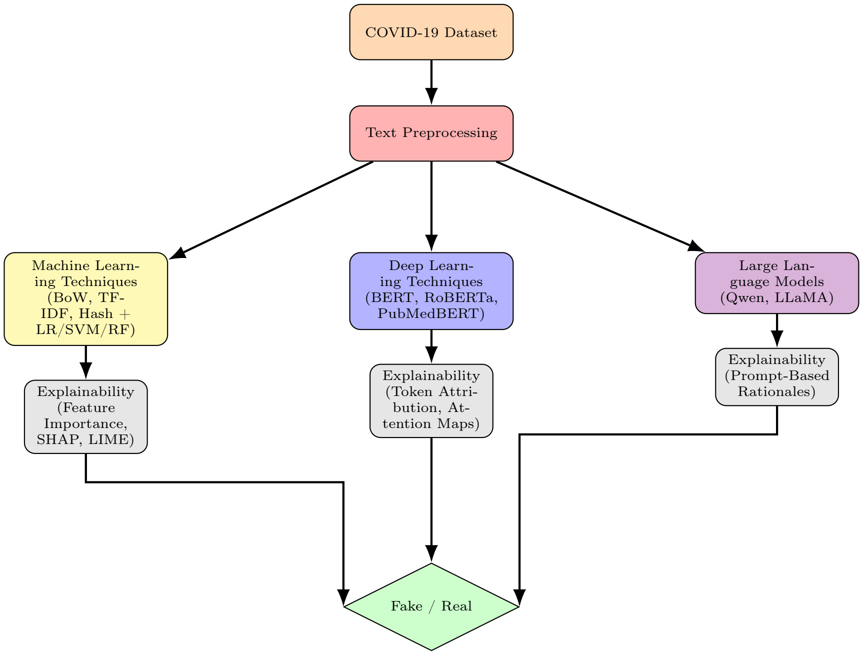
COVID-19 Dataset
Text Preprocessing
Machine Learning Techniques (BoW, TF-IDF, Hash + LR/SVM/RF)
Deep Learning Techniques (BERT, RoBERTa, PubMedBERT)
Large Language Models (Qwen, LLaMA)
Explainability (Feature Importance, SHAP, LIME)
Explainability (Token Attribution, Attention Maps)
Explainability (Prompt-Based Rationales)
Fake / Real

Fig. 1. Framework showing Machine Learning, Deep Learning, and Large Language Model pipelines with model-specific explainability leading to final fake/real classification.

text. Transformer architectures such as BERT and MedBERT are fine-tuned end-to-end on the fake news detection task, allowing both the embedding layers and classification heads to be jointly optimized. This stage enables the modeling of long-range dependencies, contextual meaning, and domain-specific language patterns, particularly relevant for healthcare misinformation.

In the final stage, lightweight large language models are integrated into the framework to evaluate reasoning-driven fake news detection. Models such as Qwen 2.5 3B and LLaMA 3.2 3B are employed to perform classification by assessing global narrative coherence and contextual plausibility. These models operate at a higher level of abstraction compared to earlier approaches, enabling robust detection of subtle and semantically complex misinformation.

To ensure transparency and trustworthiness, explainable artificial intelligence techniques are applied across all stages of the framework. Feature attribution methods are used for word frequency based models, token-level and contextual explanations are applied to transformer models, and prompt-based rationale generation and perturbation analysis are employed for large language models. The outputs from all model families are evaluated using identical performance metrics, including Accuracy, Precision, Recall, F1-score, and AUC, to enable fair and consistent comparison across different representation learning paradigms.

4.2 Text Preprocessing

Text preprocessing includes lowercasing, noise removal, stopword elimination, tokenization, and morphological normalization using stemming and lemmatization to reduce sparsity while preserving semantics.

4.3 Text Encoding

In the first stage, frequency-based representations are employed to capture surface-level lexical patterns in text.

- Count Vectorizer represents documents as sparse vectors of term frequencies.
- TF-IDF weights terms based on their importance within individual documents and across the corpus.
- HashingVectorizer maps tokens to fixed-length feature vectors using a hashing function, enabling memory-efficient encoding without vocabulary storage.

These representations are combined with classical machine learning classifiers, including Logistic Regression, Support Vector Machines, Random Forests, and Gradient Boosting, to establish strong and interpretable baseline performance.

4.4 Word Embeddings

To capture semantic relationships beyond lexical frequency, the framework incorporates both static and contextual word embeddings.

Static Word Embeddings Pretrained static embeddings, namely Word2Vec, GloVe, and FastText, are used to generate dense vector representations of words learned from large unlabeled corpora. These embeddings encode syntactic and semantic similarities and are supplied as input to downstream classifiers. FastText further improves robustness by modeling subword information, enabling effective handling of out-of-vocabulary terms. Despite their advantages, static embeddings assign a single representation per word and therefore lack contextual sensitivity.

Contextual Embeddings Contextual embeddings are generated using transformer-based architectures such as BERT and RoBERTa, along with domain-specific models including MedBERT. These models employ self-attention mechanisms to generate dynamic token representations conditioned on surrounding context, allowing effective modeling of long-range dependencies and healthcare-specific language patterns. The models are fine-tuned end-to-end on the fake news detection task to jointly optimize representation learning and classification.

4.5 Large Language Models

Large Language Models (LLMs) extend transformer-based architectures by scaling the number of parameters and training data, enabling reasoning over long contexts and complex semantic relationships. Given an input text sequence $X = \{x_1, x_2, \ldots, x_n\}$, LLMs model the conditional probability of the next token using an autoregressive formulation:

$$P(x_t \mid x_1, x_2, \ldots, x_{t-1}) = \mathrm{Softmax}(W h_{t-1}) \tag{1}$$

where h_{t-1} represents the hidden state produced by stacked self-attention and feed-forward layers, and W is the output projection matrix. The self-attention mechanism computes contextual representations as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{2}$$

allowing each token to attend to all other tokens in the sequence.

In this study, lightweight LLMs such as Qwen 2.5 3B and LLaMA 3.2 3B are employed to perform fake news detection through instruction-based prompting. Instead of learning a task-specific classifier, the model is guided to generate a label by reasoning over the full input text. The classification decision is derived from the generated output:

$$\hat{y} = \arg \max_{c \in \{\mathrm{Fake}, \mathrm{Real}\}} P(c \mid X) \tag{3}$$

This formulation enables LLMs to evaluate global narrative consistency and semantic plausibility rather than relying solely on localized token patterns, making them effective for detecting subtle and context-dependent misinformation.

4.6 Explainability Pipeline

The explainability pipeline is designed to provide model-specific interpretability across different stages of the framework. For large language models, traditional gradient-based or feature attribution methods are not directly applicable due to their generative and autoregressive nature. Therefore, explainability is achieved using behavior-based interpretation strategies.

Prompt-Based Rationale Generation In prompt-based explainability, the input text X is augmented with an instruction requesting both a prediction and an explanation. The model generates a rationale R conditioned on the input:

$$P(R \mid X) = \prod_{t=1}^{|R|} P(r_t \mid X, r_1, \ldots, r_{t-1}) \tag{4}$$

The generated rationale highlights the semantic cues and reasoning patterns that contribute to the predicted label, providing an interpretable explanation aligned with the model's internal reasoning process.

5 Experimental Settings

5.1 Experimental Setup

All experiments were conducted on Google Colaboratory using Python-based libraries, including scikit-learn, TensorFlow, NumPy, and pandas. GPU acceleration was used for transformer and large language model experiments.

5.2 Dataset

The COVID-19 Fake News Detection dataset, introduced as part of the shared task `CONSTRAINT@AAAI-2021`, contains 10,700 English social media posts and news articles related to the COVID-19 pandemic. Each entry is manually annotated as either *Fake* or *Real*, supporting research in misinformation detection. The dataset was curated from various public online sources, ensuring a balanced distribution between the two classes to facilitate supervised learning.

Composition & Splits.

- Training set: 6,418 instances
- Validation set: 2,134 instances
- Test set: 2,148 instances (labels withheld for competition evaluation)
- Labels: Fake, Real
- Language: English

Table 1. Distribution of fake and real news in the COVID-19 FND dataset

Dataset Name	**Fake News**	**Real News**	**Total**
COVID-19 Dataset	5,100	5,600	10,700

5.3 Evaluation Metrics

To assess the performance of the classification models, a standard set of evaluation metrics was employed. These metrics are derived from the confusion matrix, which tabulates the model's predictions against the actual labels. The core components of this matrix are:

- **True Positives (TP):** The number of instances correctly predicted as positive (e.g., fake news correctly identified as fake).
- **True Negatives (TN):** The number of instances correctly predicted as negative (e.g., real news correctly identified as real).
- **False Positives (FP):** The number of instances incorrectly predicted as positive (e.g., real news incorrectly identified as fake). Also known as a Type I error.
- **False Negatives (FN):** The number of instances incorrectly predicted as negative (e.g., fake news incorrectly identified as real). Also known as a Type II error.

Based on these components, the following metrics were calculated:

Accuracy: This metric measures the proportion of all predictions that were correct. It is a good general measure but can be misleading on imbalanced datasets.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Precision: This metric, also known as positive predictive value, measures the proportion of positive predictions that were actually correct. It answers the question: "Of all the news we labeled as fake, how many were actually fake?"

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Recall: Also known as sensitivity or the true positive rate, this metric measures the proportion of actual positives that were correctly identified. It answers the question: "Of all the actual fake news, how many did we successfully find?"

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

F1-Score: This is the harmonic mean of Precision and Recall. It provides a single, balanced score, which is particularly useful when there is an uneven class distribution.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

AUC (Area Under the ROC Curve): The Area Under the Receiver Operating Characteristic (ROC) Curve is a performance measurement for classification problems at various threshold settings. It measures the model's ability to distinguish between classes. An AUC of 1.0 indicates a perfect classifier, while an AUC of 0.5 indicates performance no better than random chance.

6 Results and Analysis

This section summarizes model performance across representation paradigms, with quantitative results reported in Tables 2–5.

6.1 Traditional Embeddings + ML Algorithms

Here, we present the baseline performance of traditional machine learning classifiers when combined with three common frequency-based text vectorization techniques. The results are summarized in Table 2.

Table 2. Combined Accuracy and F1 Scores for Text Feature Extraction Methods

Vectorizer	**Classifier**	**Unprocessed**		**Stemming**		**Lemmatisation**	
		Accuracy	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
BAG OF WORDS	Random Forest	92.29	0.9228	93.64	0.9363	93.55	0.9354
	SVC	93.32	0.9330	93.88	0.9387	93.55	0.9354
	Logistic Regression	**94.58**	**0.9457**	**94.21**	**0.9420**	**93.93**	**0.9392**
TF-IDF	Random Forest	92.15	0.9214	93.13	0.9312	93.22	0.9321
	SVC	**94.25**	**0.9424**	**94.35**	**0.9433**	**94.30**	**0.9429**
	Logistic Regression	92.48	0.9245	93.04	0.9302	92.94	0.9293
HASH VECTORIZER	Random Forest	92.01	0.9199	93.08	0.9307	93.50	0.9349
	SVC	**94.30**	**0.9429**	**94.02**	**0.9401**	**94.39**	**0.9438**
	Logistic Regression	91.26	0.9124	91.45	0.9143	91.68	0.9167

6.2 Static Word Embeddings + ML Algorithms

Here, we present the performance of traditional machine learning classifiers when combined with three widely used static word embedding techniques, namely GloVe, Word2Vec, and FastText. These embeddings capture semantic and syntactic relationships between words by leveraging large unlabeled corpora, thus providing richer representations than frequency-based methods. The experimental results, summarised in Table 3, demonstrate how different classifiers perform across unprocessed, stemmed, and lemmatized text, highlighting the impact of preprocessing on embedding-based approaches.

Table 3. Combined Accuracy and F1-Scores for Word Embedding Methods

Word Embedding	**Classifier**	**Unprocessed**		**Stemming**		**Lemmatisation**	
		Accuracy	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
GloVe	Random Forest	90.18	0.9015	88.27	0.8824	89.85	0.8982
	SVC	**90.88**	**0.9083**	**90.56**	**0.9050**	**91.82**	**0.9177**
	Logistic Regression	90.42	0.9036	85.93	0.8582	88.27	0.8819
Word2Vec	Random Forest	90.84	0.9081	87.19	0.8717	89.39	0.8937
	SVC	**93.17**	**0.9314**	**91.44**	**0.9140**	**92.75**	**0.9272**
	Logistic Regression	90.46	0.8999	87.28	0.8720	89.90	0.8984
FastText	Random Forest	90.84	0.9079	87.75	0.8770	89.90	0.8986
	SVC	**91.02**	**0.9096**	**90.37**	**0.9030**	**91.63**	**0.9157**
	Gradient Boost	90.88	0.9083	87.47	0.8738	89.01	0.8894

6.3 Transformer Based Models

Finally, we assess the performance of modern transformer-based models, which are known for their ability to understand deep contextual relationships in text. Table 4 details the accuracy of BERT and RoBERTa base models.

Table 4. Performance (Accuracy %) of Transformer-Based Models

Model	**Accuracy (%)**
BERT	97.33
RoBERTa	95.51
MedBERT	97.43

The strong performance of PubMedBERT highlights the benefits of domain-specific pretraining for health-related misinformation detection.

6.4 Large Language Model Performance

Table 5 presents the performance of large language models evaluated in a zero-shot setting.

Table 5. Performance of Large Language Models (Zero-Shot)

Model	**Accuracy (%)**
LLaMA-2 (7B)	97.80
Qwen-7B-Instruct	92.36

These results suggest that LLMs offer strong generalization and reasoning capabilities, particularly in settings where labeled data are scarce.

7 Explainability Analysis

To analyze the decision-making behavior of different fake news detection models, explainable artificial intelligence techniques are applied across all model families. The objective of this analysis is to understand how predictions are formed at different representation levels, ranging from lexical feature reliance to contextual understanding and reasoning-driven decisions. The explainability results are used to assess transparency, identify biases, and evaluate the trustworthiness of the proposed framework.

7.1 Lexical Feature Attribution (XAI1)

Lexical feature attribution is applied to traditional machine learning models trained using word frequency based representations such as TF-IDF, CountVectorizer, and HashingVectorizer. Feature importance scores derived from linear classifiers indicate that predictions are primarily driven by the presence or absence of specific keywords. Terms commonly associated with misinformation narratives, sensational expressions, and repeated domain-specific keywords contribute positively toward fake predictions, while institutional terminology and formal reporting language are more strongly associated with real news classification.

The analysis highlights that lexical models rely heavily on surface-level word frequency patterns. While this behavior provides clear interpretability, it also exposes sensitivity to lexical bias, where the presence of certain trigger words can disproportionately influence predictions regardless of broader semantic context.

7.2 Contextual Token Attribution (XAI2)

Contextual token attribution is performed on transformer-based models to examine how context-aware representations influence classification decisions. The explanations show that predictions are shaped by interactions between tokens and their surrounding context rather than isolated keywords. Influential tokens often appear within meaningful phrases that convey claims, uncertainty, or authoritative statements.

Compared to lexical models, transformer-based explanations demonstrate reduced dependence on individual trigger words and greater emphasis on sentence structure and contextual relationships. This shift enables improved handling of ambiguous and nuanced news content, reflecting a move from keyword matching toward contextual understanding.

7.3 Semantic Contribution Analysis (XAI3)

Semantic contribution analysis is conducted to assess how higher-level semantic information affects model predictions. The explanations reveal that model decisions are influenced by distributed semantic cues spanning multiple related words and phrases rather than single tokens. This indicates that predictions are based on overall narrative coherence and thematic consistency.

The results show that semantic representations help distinguish between misleading and factual content even when there is substantial lexical overlap. This level of explainability provides a more holistic interpretation of model behavior, bridging the gap between contextual token-based models and reasoning-driven large language models.

7.4 Explainability of LLM-based Models (XAI4)

Explainability for large language models is addressed exclusively through prompt-based rationale generation. In this approach, the model is instructed to produce both a classification label and a concise explanation supporting its decision. The generated rationales typically reference narrative consistency, plausibility of claims, and alignment with known real-world information rather than individual words or tokens.

The rationale-based explanations suggest that large language models rely on reasoning over the full input text, evaluating semantic coherence and factual plausibility at a global level. However, as these explanations are generated outputs rather than direct feature attributions, they are interpreted as indicative of the model's reasoning behavior rather than guaranteed faithful representations of internal computations. This limitation is explicitly acknowledged, and prompt-based rationales are used to provide qualitative insight into LLM decision-making.

8 Conclusion and Future Work

This paper presented an explainable fake news detection framework that evaluates word frequency based models, static word embeddings, contextual transformer architectures, and lightweight large language models within a unified setting. The results demonstrate a clear progression in detection capability from lexical feature reliance to contextual understanding and reasoning-driven predictions, with large language models achieving the highest performance. The explainability analysis highlighted distinct decision-making behaviors across model families. Lexical models depend on keyword frequency, transformer-based models emphasize contextual token interactions, and large language models reason at a global semantic and narrative level. Prompt-based rationale generation provided qualitative insights into the reasoning process of large language models, supporting interpretability despite their inherent opacity. Future work will focus on improving the reliability of explainability for large language models, extending the framework to multimodal misinformation detection, and evaluating larger and more diverse datasets to enhance robustness and real-world applicability.

References

1. Cao, J., Sheng, Q., He, J., Li, J.: SLIM: A systematically selected, limited information-based method for fake news detection. In: *Proceedings of the IEEE International Conference on Big Data*, Osaka, Japan, pp. 1380–1389 (2022)
2. Chen, T., Fu, Z., McKeown, K.: Real-time factuality assessment from adversarial feedback. *arXiv preprint arXiv:2405.15835* (2024)
3. Hu, S., Chen, Z., Wu, T., Lee, L.H., Hooi, B.: Bad actor, good advisor: Exploring the role of large language models in fake news detection. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, pp. 1152–1166 (2023)
4. Islam, M.R., Liu, J., Wang, X.: LUND-FN: A large benchmark fake news detection dataset for the Urdu language. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Dubrovnik, Croatia, pp. 2503–2516 (2023)
5. Li, Y., Li, Q., Li, H., Liu, Y., Gao, K.: Fake news detection through topic and sentiment analysis based on fusion of graph convolutional network and BERT. *Applied Soft Computing* **111**, 107693 (2021)
6. Mukherjee, S., Ghosh, S.: UNITE-FND: Unimodal text-driven multimodal fake news detection. *arXiv preprint arXiv:2405.19530* (2024)
7. Patwa, P., et al.: Fighting an infodemic: COVID-19 fake news dataset. In: *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations (CONSTRAINT)*, pp. 21–31 (2021)
8. Su, Y., Shu, K., Liu, H.: How to adapt your fake news detector to the era of LLMs? *arXiv preprint arXiv:2305.18739* (2023)

9. Yan, Y., Zheng, L., Zhang, X., Zhang, Y.: BTCM: A BERT-based 1D-CCNet attention model for multimodal fake news detection. *Electronics* **12**(14), 3045 (2023)
10. Xu, S., Li, Y.: A comparative study of online and offline large language models for fake news detection. *arXiv preprint arXiv:2405.08600* (2024)
11. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* **3**(1), 1–23 (2021)
12. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017)
13. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. In: *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144 (2016)
14. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774 (2017)