VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2026, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 14, Number 1, January-March 2026 *pp:1-18*

Explainable BERT Models for Rumor Detection on Social Media

Aziz Ahmad[1], **Shams Ur Rahman**[1*], **Spogmay Yousafzai**[1], **Ghulam Hafeez**[2]

[1]Department of Computer Software Engineering, University of Engineering & Technology, Mardan, Pakistan;
[2]Department of Electrical Engineering, University of Engineering & Technology, Mardan, Pakistan

Keywords: *Rumor detection, Social media, Transformers, BERT, Retrieval-Augmented Generation, Explainable AI, LIME.*

Journal Info:
Submitted:
November 04, 2025
Accepted:
February 05, 2026
Published:
February 13, 2026

Abstract The rapid spread of unverified information and rumors across social media platforms poses serious risks to public health, economic stability, and societal trust. Researchers developed specific machine learning and deep learning models for automated rumor detection because of the increasing challenge that these platforms posed. The existing automated systems operate through uninterpretable black box operations thus users find it hard to trust these systems or understand how they work. This study addresses these limitations by developing explainable framework for rumor detection through transformer-based models along with advanced Large Language Models (LLMs). The proposed research evaluates and compare the performance outcomes between fine-tuned BERT-based models BERT, RoBERTa, DistilBERT, and ALBERT with cutting edge Large Language Models (LLMs) LLaMA3, DeepSeek R1, and Mistral. These models are applied to the PHEME dataset, a corpus contains actual Twitter posts that have already received labels as either rumors or unverified statements or non-rumors. Data preprocessing includes cleaning tweet text, extracting engagement metrics, and user features. BERT performs best in the fine-tuned context by achieving the highest accuracy (82.9%) and the RAG-based LLMs showed less effective in the zero-shot evaluation; thus, the observed discrepancy indicates variations in training paradigms rather than inherent architectural superiority. Explainability is achieved using Local Interpretable Model-Agnostic Explanations (LIME), which visualizes influential features behind predictions. The findings highlight a trade-off between LLM flexibility and transformer precision, offering a scalable, interpretable solution for trustworthy rumor detection and content moderation.

***Correspondence author email address:** shams@uetmardan.edu.pk
DOI: 10.21015/vtse.v14i1.2275

1 Introduction

Social media has revolutionized information dissemination, becoming main news media for millions around the world. The speed and lack of moderation at which content can spread on these platforms however has helped spread rumors and misinformation. Rumors, unverified or false information that seem credible can spread fast, and cause public panic, confusion and societal harm. A notable example occurred in 2013 when Twitter account was hacked and falsely reported an explosion at the White House to trigger a large scale stock market downturn [1]. Such events highlight the strong need for effective and reliable techniques that will detect the spread of rumors in social media. Traditional rumor detection approaches have relied on manual feature extraction and application of machine learning algorithms. Although these methods have achieved some success, they are usually time consuming and

labour intensive and not very effective in identifying the dynamic nature of data in social media [2]. The introduction of deep learning, especially transformer based model such as BERT (Bidirectional Encoder Representations from Transformers) has highly automated feature extraction and enhanced detection accuracy [3]. Despite these developments, such models are criticized for their lack of interpretability in that they function as "black boxes" not revealing much about their decision processes. This opacity makes high stakes applications such as rumor detection challenging given the need for grasping the rationale behind predictions needed.

To overcome these limitations, this study suggests a unified and explainable rumor detection framework that systematically compares fine tuned transformer based classifiers with Retrieval-Augmented Generation (RAG)-based Large Language Models. Compared to previous studies that analyze either the accuracy or the contextual reasoning separately, the given work analyzes the performance, interpretability, and reproducibility simultaneously on the background of the common dataset, standardized metrics, and explainable AI methods.

This research utilizes the state-of-the-art transformer based architectures, including BERT [3], RoBERTa [4], ALBERT [5] and DistilBERT [6], as well as advanced Large Language Models (LLMs) such as LLaMA3 [7], DeepSeek R1 [8], and Mistral [9]. The transformer-based models in this paper are trained to discover rumor, non-rumor and unverified labels on the PHEME dataset, whereas the LLMs are tested in the zero-shot setting within a Retrieval-Augmented Generation (RAG) framework without being trained to use the PHEME labels.

Additionally, the combination between LLMs and Retrieval-Augmented Generation (RAG) allows the extraction of relevant context; hence, improving the models understanding and classification competence [10]. A key aspect of this research is the usage of the explainable AI features for the analysis of the models decision making processes. In particular, Local Interpretable Model-agnostic Explanations (LIME) [11] are used to explain how predictions are made and to detect possible biases, and to increase the interpretability of the models. By focusing on the approach of LIME, the study aims to provide transparent explanations without the complexity of other approaches such as SHAP or Integrated Gradients. The significance of this study lies in its potential to contribute to the development of more transparent and trustworthy AI systems for rumor detection. Through addressing the limitations of the existing solutions, this research aims to offer a robust and scalable solution for preventing the spread of misinformation through social media. There is therefore important implication of the findings of this study in the realm of public health, economic stability and social cohesion, and that should necessitate the continuation of research and innovation in this critical field.

2 Literature Review

The proliferation of social media platforms has revolutionized information dissemination, making them primary sources for news and updates. Nevertheless, the ease with which information can disseminate has also promoted speedy spread of rumour and misinformation, whereby leading to potential public panic and confusion. As such, the development of automated detection of rumors systems has become a necessity to ensure information integrity and public trust. Here, we explore the way these methodologies have been evolving from the wide range of traditional machine learning methods, deep learning models through to transformers-based architectures, graph- based approaches and the use of explainable AI (XAI) techniques.

2.1 Traditional Approaches to Rumor Detection

In early studies of rumor detection, there were frequent comparisons between the spread of information and the spread of infectious diseases, resulting in the application of epidemiological models to explain and forecast rumor propagation dynamics. These models provided the basis for analyzing the ways in which rumors spread over social networks. Zhu et al [12] developed SI model used to model rumor spread in which situations individuals become infected when exposed to a rumor, in turn, becoming permanently infected. Under this framework, Jiang et al. [13] introduced the idea of "rumor centrality" with the goal of detecting the most likely origin of the rumor on the network basis. Their approach is based on a regular tree topology that though mathematically tractable, it is limited in its applications in real social networks that often display more complicated struc-

tures. In order to address some of the limitations with the SI model the susceptible infected recovered (SIR) model was used, which added a recovered state to capture people who stop spreading the rumor after a certain amount of time. Malhotra et al.[14] introduced a reverse-infection algorithm for source detection, though it was limited in networks with cycles. The SEIZ model further included "exposed" and "skeptic" states to better represent user diversity. These limitations highlighted the need for more sophisticated approaches that integrate network dynamics, content analysis, and user behavior in order to improve the effectiveness of rumor detection and their source identification in social media settings.

2.2 Machine Learning and Temporal Analysis

Transitioning from theoretical models, researchers started using machine learning techniques to optimize rumor detection. Kotteti et al. [15] implemented temporal tweet features with Gaussian Naive Bayes, which showed a high precision (0.94) but low recall. To address the limitations of relying on only temporal features, researchers have pursued an interest in integrating spatial and temporal structures. For instance, Huang et al. [16] used a more holistic approach to describe the message flow using a spatial temporal structure neural network (STS-NN). Thakur et al. [17] proposed a supervised ML system based on the study of the linguistics of the tweets that performed better than 81%. Recent studies have also aimed at revealing temporal pattern in information dissemination. A research conducted by [18] reported upon textual and temporal properties of the social media text referring to coronavirus in order to enhance rumour detection determining the change of information in time.

Overall, the set of studies is indicative of the importance of introducing temporal analysis and machine learning methods to improve the efficiency and timeliness of rumor identification on social media services.

2.3 Deep Learning Models

The introduction of deep learning has been a revolutionary turn around in the field of rumor detection because the capability of computational systems to learn hierarchical representations directly from raw social media data has been greatly improved. Deep learning models do not require manual feature engineering as done with tradition machine learning methods hence can automatically extract semantic, contextual, and temporal features which can allow more accurate and scalable rumor classification. A particular contribution is that of Chen et al. [19] that proposed a multimodal model that integrates LSTM and RoBERTa model embeddings into graph-structured conversation. Ajao et al. [20] introduced the hybrid CNN and LSTM model whose approach reflected capturing the syntactic sequential features without dependency on the data specific events. Zhou et al. [21] proposed an early rumor detection (ERD) system to develop a reinforcement learning mechanism to strike the balance between the time and accuracy of rumor identification. Ma et al. [22] have offered the CallAtRumors, which employed a soft attention mechanism in RNNs, which it employed to underscore significant posts. Wang & Guo [23] have used a cascaded architecture of GRUs with encoding of sentiment and semantics, an accuracy of 88.5% on Twitter16.

Lastly, the dominance of transformer-based architectures was validated by Kaliyar et al. [24] when BERT was used for a rumor detection task. The model, which was trained on posts from social media sites, performed overwhelmingly better than the traditional machine learning baselines when using pre-trained language representations capable of expressing deep contextual dependencies. This research not only reported the resistance of BERT against noisy, small- length texts but also its tunability in various misinformation detection scenarios. Taken altogether, these studies represent the range of diverse strategies in deep learning that target priority concerns about rumor detection, such as structure modeling, temporal logic, early intervention, and semantic interpretation. However, these models, despite their outstanding performance, tend to lack interpretability, a limitation that is in need of incorporating deliberations on explainable AI technique as discussed in the following sections.

2.4 Transformer Based Models and Explainability

The significant contributions to rumor detection through the use of transformer-based architectures lie in their abilities to model intricate linguistic patterns and con-

textual relationships. Anggrainingsih et al. [25] used BERT for rumor detection, achieving 86.9% accuracy on PHEME. Despite high performance, these models act as "black boxes." In order to address the challenge of interpretability, Sharma and Sharma [26] applied a weakly supervised learning method to detect potential rumor spreaders in Twitter. By including the user behaviors, textual contents and the ego-network features, they improved the detection process with an F1-score of 0.864 and an AUC-ROC of 0.720. Their approach proved that it was possible to enhance the ability of the model to detect individuals likely to spread rumors by adding user-centric features. Pattanaik et al. [27] combined BERT and GPT embeddings with Graph Convolutional Networks (GCNs), resulting in an accuracy of 88.64% accuracy on the PHEME dataset. In spite of this ensemble approach excellent performance, the model had challenges of computational complexity and low explainability. To enhance generalizability and interpretability Joshi et al. [28] combine Domain Adversarial Neural Networks (DANN) with Local Interpretable Model-Agnostic Explanations (LIME). Their framework was designed to discover misinformation spanning over various social media by learning domain- invariant features. The addition of LIME gave post-hoc explanation of the model's prediction, which in turn increased transparency and trustworthiness of model. Khoo et al. [29] proposed the PLAN model, using multi- head attention to generate token- and post-level explanations. Its variants (StA-PLAN, StA-HiTPLAN) used hierarchical attention and structural knowledge to make it more interpretable. These developments emphasize the need to strike a balance between the performance of the model and interpretability in rumor detection. Further studies should focus on further developing hybrid models that combine the transformer architecture with other sources and explainable AI approaches to ensure that the resulting system is both accurate and transparent in detecting and preventing the spread of misinformation.

2.5 Large Language Models and Retrieval-Augmented Generation and Explianability

The advent of Large Language Models (LLMs) has significantly advanced in the field of natural language processing, as machines can generate coherent text and understand complex contexts. Notable LLMs include GPT-2 [30], LLaMA [31], DeepSeek [8], and Mistral [9]. These models have shown remarkable performance in many languages tasks. When combined with Retrieval-Augmented Generation (RAG) frameworks, the incorporation of LLMs has improved their practical capabilities, especially in knowledge-intensive tasks further. RAG brings together, the generative power of LLMs with retrieval mechanisms that introduce external knowledge during inference. With this integration, models are given access to up to date information which helps them to be effective when responding to complex queries and reduces the probability of generating outdated or wrong responses. Introducing RAG framework, Lewis et al. [10] proved it to be effective for open-domain question answering by integrating the retrieval mechanisms with generative models. Nevertheless, the use of LLMs and RAG to rumor detection is still in development. Although it is promising, it has some challenges such as a high computational cost and low interpretability because their black-box nature makes its application in sensitive contexts where explainability is essential difficult.

The above-mentioned studies reveal several limitations, including limited interpretability of deep learning and transformer-based rumor detection models, difficulties in handling short, noisy, and event-driven social media text, high computational complexity of advanced architectures, and the absence of systematic and reproducible comparisons across different model families. Although recent approaches based on Large Language Models and Retrieval-Augmented Generation demonstrate promising contextual reasoning capabilities, their effectiveness for rumor detection remains underexplored, particularly in real-time and explainability-sensitive settings. To address these challenges, a unified and explainable evaluation framework is required that jointly assesses performance and interpretability to enable reliable and transparent rumor detection on social media platforms.

3 Methodology

The proposed research methodology for developing explainable rumor detection system on social media include several important phases: data collection and data

preprocessing, training the transformer-based models and Large Language Models (LLMs), and incorporate explainable AI (XAI) technique, and evaluating model performance.

The PHEME dataset is the main source of data used which consists of real-world Twitter posts, which have been labeled as either rumor or non-rumor or unverified. BERT-based models like BERT, DistilBERT, ALBERT, RoBERTa are fine-tuned for classification. At the same time, there is the use of advanced LLMs such as LLaMA, and Mistral in a Retrieval-Augmented Generation (RAG) framework to improve understanding of context. To understand the predictions and bring transparency, explainable AI method, such as LIME are used. Figure 1 shows the entire workflow of the proposed approach, starting with data ingestion and ending with the final prediction of coefficients values.

3.1 Dataset

The PHEME dataset was selected for this study due to its realistic data driven by events and is a good benchmark of rumor detection in social media. PHEME is annotated with rumor, non-rumor, and unverified, unlike synthetic or purely text-level datasets, containing the temporal and conversational dynamics of information diffusion in the context of breaking news events. Its wide use in previous rumor detection studies also allows its meaningful comparison with current methodologies and facilitates the reproducibility and external validity of the results of the experiment.

For the proposed approach, we used the PHEME dataset for the detection of rumors within social media platforms especially twitter. The dataset consists of 2,402 of conversation threads in Twitter that are annotated as rumor, non-rumor, or unverified, and organized around nine actual events that drew much public interest. Charlie Hebdo shooting, Ottawa Parliament shooting, footballer Michael Essien's Ebola virus hoax, the hoax of the Prince concert in Toronto, Ferguson unrest, disappearance of Vladimir Putin, Germanwings plane crash, the scandal associated with Gurlitt art collection, and hostage crisis of Sydney siege [32].

Every conversation thread in the dataset starts from a source tweet, which is followed by a stream of replies, which constitutes an open-ended discussion in threaded form, which reflects how information is spread in real time on social media. These threads are annotated corresponding to the truth in the original claims, which gives a pragmatic context for analyzing the spread and evolution of information. The annotations of the dataset enable researchers to work with the initial claim and the collective discourse that accompanies it, thus making the studies on rumors propagation and verification possible [33]. Apart from the textual content of tweets, the PHEME dataset also contains a lot of metadata, which includes user information including verification status, followers count, and account age. engagement metrics which include retweet and favourite counts; as well as contextual tags that include event labels and timestamps. This thorough structure allows linguistic as well as contextual analysis and can be applied to work with any modelling technique from traditional classifiers and deep learning models to state of the art transformers and large language models [34].

Figure 2 shows the distribution of the labels in the PHEME dataset showing distribution of annotated tweets into categories of Rumor, Non-Rumor, and unverified. As demonstrated, the verified rumor claims occupy a large proportion of the dataset (44%), the unverified observations (29%) and the false rumor information (27%) occupying the remaining dataset. This neutral, but diversified assignment of labels aids in assessment of multi-class classification models and reinforces the vitality of robust systems of rumor detection on varying levels of veracity. The event-driven division segmentation of the dataset also enables cross-event generalization and the check of the transfer learning strategies. Its complex and multi-dimensional nature makes it a perfect basis for training and evaluation of explainable AI models in dynamic environments of social media.

The PHEME data set used in this research paper is publicly available and can be found at: https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619

3.2 Data Preprocessing

Before model training, the raw Twitter data underwent comprehensive preprocessing to convert it into a structured format suitable for rumor detection. Non-informative metadata fields such as usernames, account dates, URLs, and protected status were removed due to low variance, while relevant features (verified, fav_count,

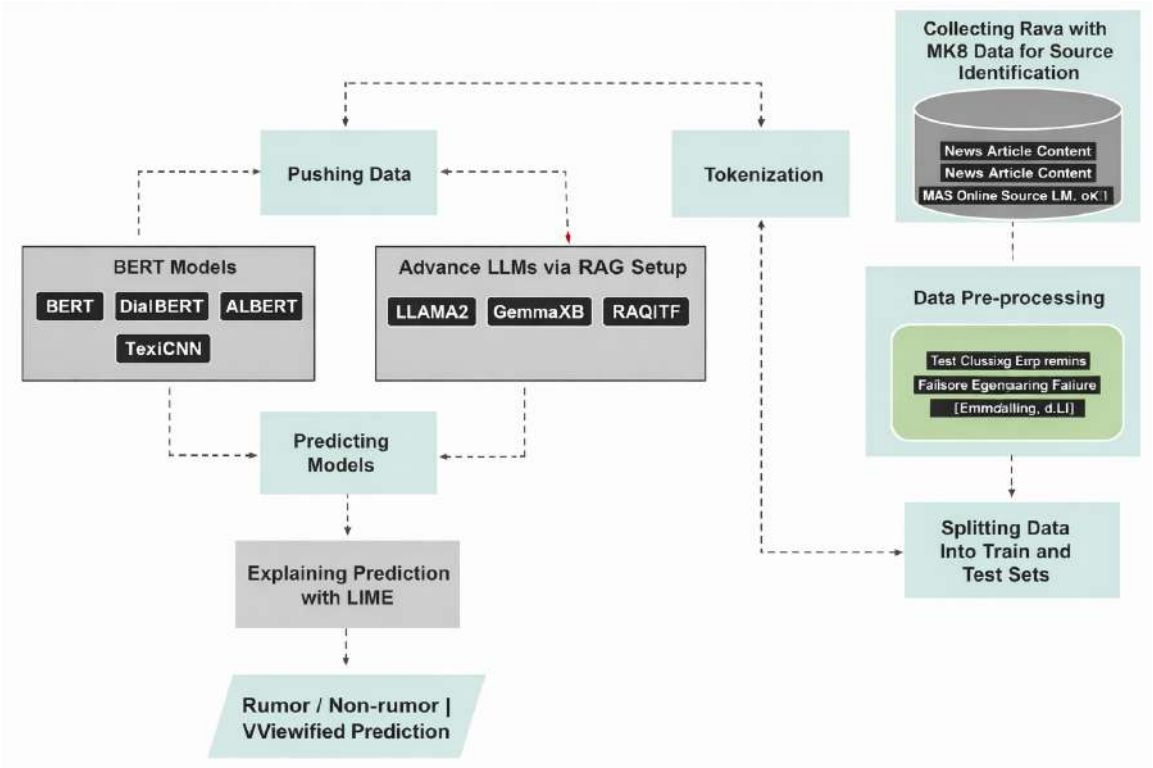

Figure 1. Proposed Research Methodology

retweet_count, followers, followings, and tweet_count) were retained and standardized. Tweets were normalized by lowercasing, removing special characters and hyperlinks via regular expressions, and eliminating English stopwords using the NLTK library. Tokenization prepared the text for transformer inputs. Lemmatization and stemming were evaluated but excluded due to minimal benefit. Rumor labels were encoded as integers (0: unverified, 1: rumor, 2: non-rumor), and selected numerical features (verified, fav_count, retweet_count, followers, followings, and tweet_count) were scaled to support model convergence and interpretability. The dataset was split using a stratified 80:20 train-test ratio to ensure class balance, enabling consistent evaluation across BERT, RoBERTa, DistilBERT, ALBERT, and RAG-based LLM pipelines. To prepare inputs for transformer-based and RAG-based large language models, tokenization and padding were applied using model-specific tokenizers from the Hugging Face library. BERT, RoBERTa, DistilBERT, and ALBERT each utilized their corresponding tokenizer, converting tweets into subword tokens with truncation and padding (max sequence length: 128) to ensure consistency and computational efficiency.

Attention masks were generated to differentiate actual content from padding. Each tokenized sample was combined with numerical features to form multimodal inputs, wrapped into custom PyTorch datasets for efficient batching. For RAG-based LLMs (`LLaMA3`, `DeepSeek-R1`, `Mistral`), tokenization was integrated with retrieval. Tweet threads were chunked, embedded, and indexed via FAISS. Retrieved segments were re-tokenized to form structured prompts for final prediction and explanation. This two-stage process ensures both input precision and rich contextual understanding.

3.3 Proposed BERT-Based and RAG-LLM Approaches

The proposed methodology integrates two complementary components: (1) fine-tuning BERT-based transformer models with auxiliary metadata for rumor classification, and (2) leveraging advanced Large Language Models (LLMs) in a Retrieval-Augmented Generation (RAG) framework for contextual detection and explainable reasoning.

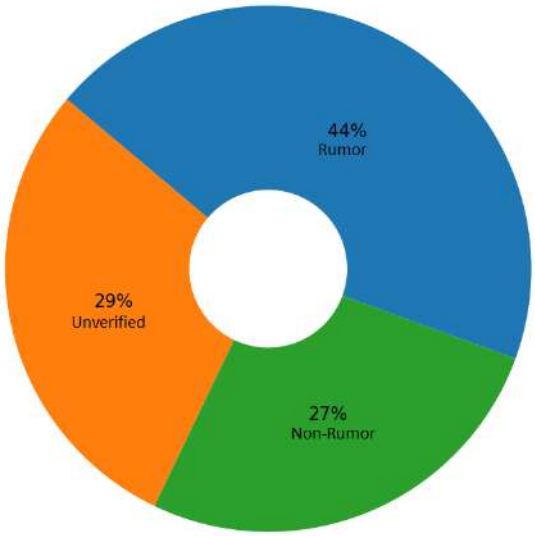

Figure 2. Label distribution of the PHEME dataset showing the percentage of tweets categorized as rumor, non-rumor, and unverified.

3.3.1 BERT-Based Transformer Models for Rumor Detection

Four transformer models (BERT, RoBERTa, DistilBERT, and ALBERT) were fine-tuned to classify tweets into three categories: rumor, non-rumor, and unverified. Every model was enhanced by structured metadata such as verified, fav_count, retweet_count, followers, followings, and tweet_count, was trained with an architecture that would utilize both text embeddings and auxiliary features in a joint manner.

$$P(y/x) = \text{Softmax}\left(W_2\ \text{ReLU}\left(W_1 \begin{bmatrix} x_{[CLS]} \\ x_{\text{aux}} \end{bmatrix} + b_1\right) + b_2\right) \quad (1)$$

The equation (1) describe forward propagation for each model architecture involves the use of [CLS] token representation output from the corresponding transformer model and concatenates it with the structured extra features (six in total). This fused representation goes through a fully connected layer with ReLU activation and dropout, and is, ultimately, classified with a softmax output over three classes. Where $h_{[CLS]}$ represents the transformer output for the [CLS] token, x_{aux} is the auxiliary feature vector, and W_1, W_2, b_1, b_2 are trainable parameters. Fine-tuning was achieved using PyTorch and the Hugging Face Transformer toolkit so that it can be accelerated via GPU. Models were optimized with AdamW and trained within 4–6 epochs with batch size 16 and learning rate 2×10^{-5}. The loss function of the choice was CrossEntropyLoss. Explainability is achieved using LIME (Local Interpretable Model-agnostic Explanations), allowing the visualization of word-level importance scores for individual predictions.

3.4 Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs)

To complement the discriminative transformers, this research integrate `LLaMA3`, `DeepSeek-R1`, and `Mistral` into a Retrieval-Augmented Generation (RAG) architecture. This hybrid framework enables explainable classification as by incorporating retrieved evidence with the generative capacity of LLMS [35]. The architecture includes the following components:

Step 1: Document Retrieval

FAISS-based retriever indexes the content of tweets with MiniLM-L6-v2 embeddings [36]. With a user query (a tweet and metadata), the top-k ($k = 5$) semantically similar documents are returned.

Step 2: Query Augmentation

Documents that were retrieved added value to the original tweet and create a broader context. This context that consists of raw text and structured metadata (followers, retweet count, etc.) is used to generate a prompt, which is used to pass to the LLM.

Step 3: Response Generation and Classification

The LLM processes the augmented input and outputs class probabilities (False Rumor, True News, Unverified), that sum to unity. It also produces natural language explanations for enhancing interpretability [37]. By using the retrieval as well as generative abilities of LLMs, the RAG approach provides transparent classification based on factual references that enables efficient detection and explanation of rumors.

Template and Probability Extraction

For RAG-based rumor detection, Large Language Models are used with the structured query prompt that combines text of the tweet and the additional metadata and retrieves available contextual evidence. The prompt was designed to enforce a fixed output format, enabling reliable probability extraction and downstream evaluation. The template used for all RAG-based LLMs (`DeepSeek-R1`, `LLaMA3`, and `Mistral`) is shown below:

Prompt Template

Tweet: {tweet_text}
Verified: {verified}, **Favorites:** {fav_count}, **Retweets:** {retweet_count},**Followers:** {followers}, **Followings:**{followings}, **Tweet Count:** {tweet_count}

Task: Classify the tweet into one of the following categories:

- False Rumor
- True News (Non-Rumor)
- Unverified

Response Format:

False Rumor: <probability between 0.0 and 1.0> True News: <probability between 0.0 and 1.0> Unverified: <probability between 0.0 and 1.0>

The documents that were retrieved based on the FAISS retriever were implicitly appended to the prompt context before the model applied inference, allowing the model to condition its prediction on both the input tweet and semantically related background information.

The regular expressions used to extract the three class-wise probability values were applied to the textual output produced by the LLM. In case the probabilities that were extracted were not equal to 1.0, the probabilities were normalized by dividing each probability by the total sum of all probabilities. Where the probability extraction had failed or the probability mass was the same, a uniform fallback distribution (1/3, 1/3, 1/3) was used. Prediction in the final classes was obtained by the use of the argmax with respect to the normalized probability vector. To be consistent with the dataset annotation scheme, False Rumor is considered to be Rumor, True News is Non-Rumor and Unverified is the same value.

Paradigm Note:
In this study, Transformer-based models trained with supervised fine-tuning on labeled rumor data of the PHEME dataset, whereas the Large Language Models (LLMs) are evaluated without supervised fine-tuning on the dataset labels in a Retrieval-Augmented Generation (RAG) framework. As a result, the reported findings are to be viewed as a comparison of fine-tuned discriminative classifier and zero-shot RAG-based reasoning baselines, instead of directly comparing architectures to architectures.

3.5 Explainable AI Technique

To enhance interpretability, this study integrates Explainable AI (XAI) using the LIME (Local Interpretable Model-Agnostic Explanations) framework across both transformer-based models and RAG-based LLMs. Given the application of rumor detection in sensitive domains such as health, politics, and crisis response, transparency in decision-making is essential.

3.5.1 LIME for Transformer Models

LIME functions based on the local approximation of the model behavior using an interpretable surrogate model, a linear regressor in most cases. LIME perturbs the input text for each prediction by the BERT, RoBERTa, DistilBERT, and ALBERT classifiers and thus evaluates the model response on the perturbed instances. It then finds out which words contribute to the computed label most. This is especially helpful for the interpretation of the effect of particular tokens or characteristics in multi-head attention-based architectures. For this implementation, a custom wrapper is built to fit transformer models to LIME's expected interface. Placeholder features are used in order to adapt to the fixed-size vector format of LIME. The model is then queried via the `predict_proba()` method and explanations are visualized as highlighted text denoting either positive or negative contributions to the final class probabilities.

3.5.2 LIME in Retrieval-Augmented LLMs

For models incorporated into the RAG framework such as `LLaMA3`, `DeepSeek-R1`, and `Mistral`, a custom wrapper is built in order to communicate with the LIME explainer.

Since LLMs work on context-augmented sequences, the input tweet is concatenated with retrieved documents from the FAISS indexer. The wrapper resolves in-context prompts and converts the LLM's output to a probabilistic form that is helpful in LIME's explanation generation. Since it is usually that LLMs output natural language instead of raw logits, a custom probability extraction logic is applied with the help of regex and normalization for compatibility. Highlighted by LIME explainer function, the most critical parts of the original tweet that led the model to make a classification decision become vivid, thus providing interpretability even in a zero-shot and few-shot classification setup.

3.6 Experimental Settings and Hyperparameters

The experiments were conducted with a standard experimental setup to guarantee reproducibility and reasonable comparison between all the models under evaluation, both transformer-based classifiers and Retrieval-Augmented Generation (RAG)-based Large Language Models. The same dataset partitions, preprocessing pipeline and evaluation metrics were used to train and evaluate all models.

Table 1 summarizes the settings and hyperparameters used in the experiment on all transformer-based and RAG-based models to be consistent in training and comparative and reproducible to the reported results.

4 Results and Discussion

The evaluation of transformer-based and retrieval-augmented generation (RAG) models has been done for rumor detection via social media, while the PHEME dataset is used for evaluation. The models were evaluated by the standard performance metrics: accuracy, precision, recall and F1-score, as shown in Section 4.1. The evaluation includes two categories of experiments: (1) Transformer-based models such as BERT, DistilBERT, RoBERTa, and ALBERT; (2) RAG-based models, including LLaMA3, DeepSeek R1, and Mistral.

4.1 Transformer-Based Models

The first experiment aimed to determine the performance of transformer models for a three-class classification problem (Rumor, Non-Rumor, Unverified). On this, a labeled dataset of 481 tweets was used, and 90% used for training, 10% for validation. Each model was trained at 5 epochs, in which auxiliary features like verification status, retweet counts, and follower counts were used. Among the transformer models, BERT showed a better performance with an accuracy of 82.95%, precision of 82.25%, Recall of 82.93% and an F1-score of 82.48%.

The confusion matrices of the evaluated models are given in Table 2 with the diagonal components reflecting the accurate predictions and non-diagonal components representing patterns of the misclassification through classes.

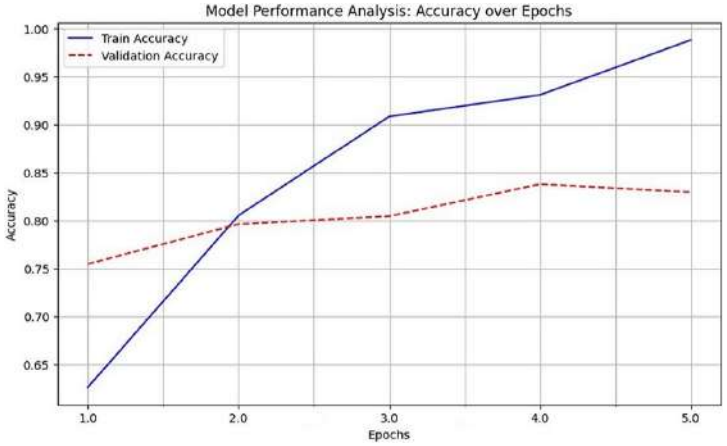

Figure 3. Training and validation accuracy of the BERT model.

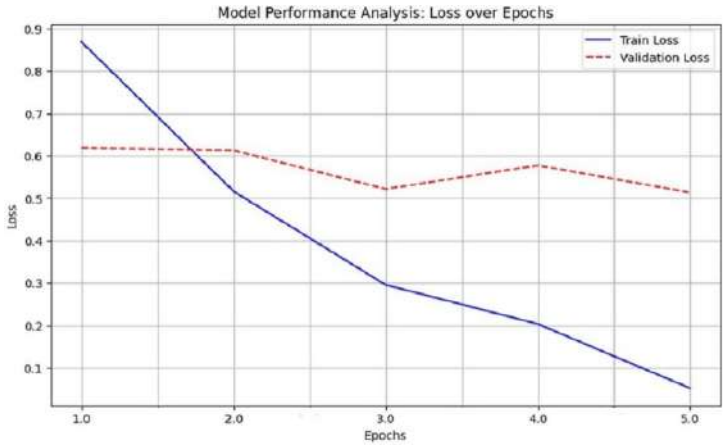

Figure 4. BERT training and validation loss.

Figure 3 shows the accuracy progression across epochs for BERT, with clear performance gains through training. Figure 4 illustrates the corresponding decline in loss, reflecting efficient learning. Similar trends were

Table 1. The key experimental settings and hyperparameters used throughout this study.

Parameter	Value
Dataset	PHEME
Train/Test Split	80% / 20% (stratified)
Classes	Rumor, Non-Rumor, Unverified
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	4–6
Max Sequence Length	128
Transformer Models	BERT, RoBERTa, DistilBERT, ALBERT
LLMs (RAG)	LLaMA3, DeepSeek-R1, Mistral
Retriever	FAISS
Top-k Retrieval	5
Explainability	LIME

observed for other transformer models.

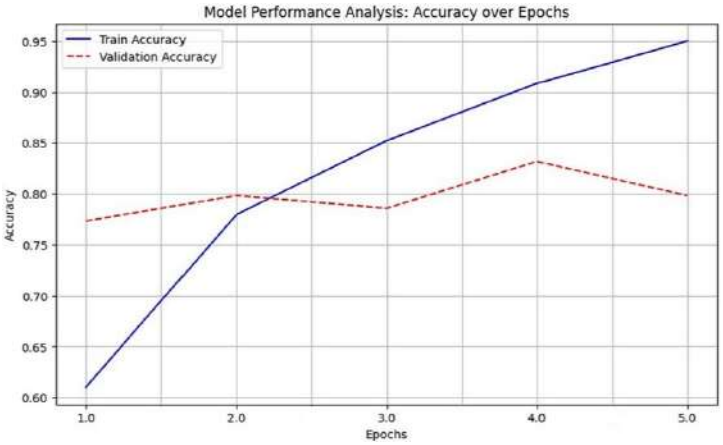

Figure 5. Training and validation accuracy of the DistilBERT model.

DistilBERT performed an overall accuracy of 80.67% (see Table 3) closely trailing BERT. Figure 5 and Figure 6 depict the accuracy and loss of the model in the course of training, remaining stable, although with a higher variance in validation's performance.

RoBERTa with a competitive performance with 81.50% accuracy as shown in Figure 7 and Figure 8. These figures show strong generalization but marginally

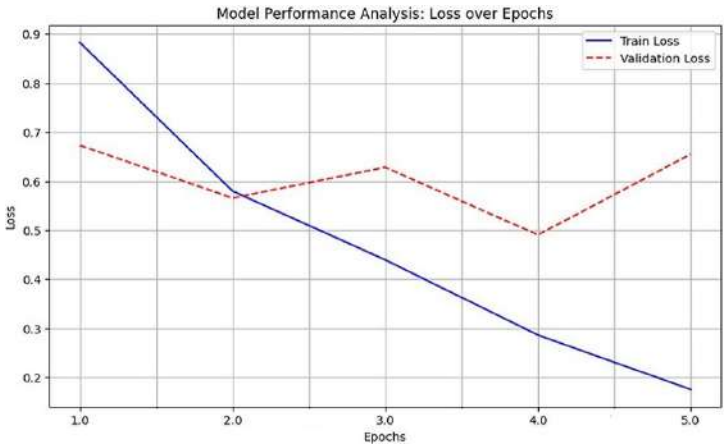

Figure 6. Training and validation loss of the DistilBERT model.

below BERT.

ALBERT matched DistilBERT in accuracy (80.67%). Figure 9 and Figure 10 show consistent learning curves, with a slight uptick in validation loss in later epochs, suggesting the need for early stopping to prevent overfitting.

The obtained results mentioned in Table 3 and the visual performance of Transformer-Based Models for Rumor Detection shown in Figure 11 are aligning with the previous studies which also show that transformer ar-

Table 2. Confusion Matrices for Transformer-Based Models.

Model	Actual Class	Rumor	Non-Rumor	Unverified
BERT	Rumor	108	12	14
	Non-Rumor	15	176	21
	Unverified	10	10	115
RoBERTa	Rumor	107	19	8
	Non-Rumor	13	179	20
	Unverified	10	19	106
DistilBERT	Rumor	114	24	6
	Non-Rumor	10	186	23
	Unverified	5	21	92
ALBERT	Rumor	114	13	7
	Non-Rumor	16	183	13
	Unverified	16	28	91

Table 3. Detection Results (%) Using Transformer Models

Sr. No	Model	Accuracy	Precision	Recall	F1 Score
1	BERT	82.95	82.25	82.93	82.48
2	RoBERTa	81.50	81.30	80.93	81.11
3	DistilBERT	80.67	80.59	79.60	79.79
4	ALBERT	80.67	80.59	79.60	79.79

chitectures are reliable in misinformation detection [38]. Fine-tuned BERT variants, with the contextual and auxiliary features, proves to have solid capabilities in handling noisy, real-world social media text [39].

4.2 Explainability with LIME (Transformer Models)

To measure the degree of interpretability, Local Interpretable Model-agnostic Explanations (LIME) was used to test samples. LIME visualizations show the significance of words regarding decision-making.

BERT Example (Figure 12): The model classified a NASA Mars rover news as Non-Rumor with 88% confidence. Words like "NASA," "lands," "rover," and "marking" contributed positively.

DistilBERT Example (Figure 13): Identified the same tweet as Non-Rumor with 42% confidence; influenced by terms like "rover", "historic", "achievement" and "exploration".

RoBERTa Example (Figure 14): The model classified the tweet as Rumor with 81% confidence, primarily influenced by tokens such as "UFO" and "spotted".

ALBERT Example (Figure 15): ALBERT classified the tweet as Rumor with 52% probability, diverging in token attribution.

4.3 RAG-Based Models

The second experiment evaluated RAG-based models on a PHEME dataset, despite computational limitations. Deepeek R1 emerged as the leading RAG model, with the accuracy of 35.0% and F1-score of 29.0%. LLaMA3 and Mistral were next with 30.0% and 25.0% of accuracy. Table 4 and Figure 16 summarizes the evaluation results

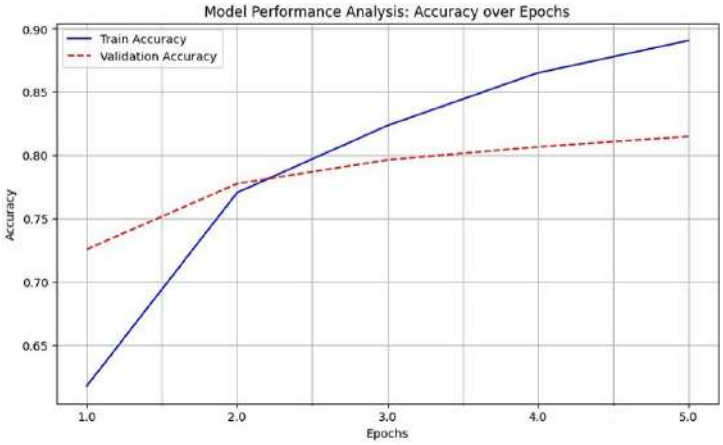

Figure 7. Accuracy plot for RoBERTa.

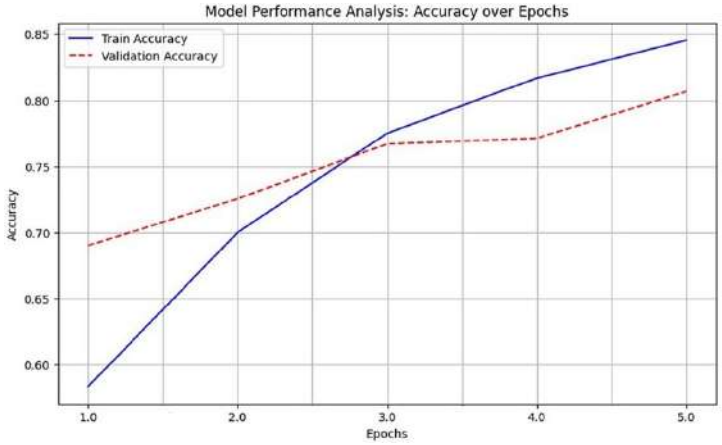

Figure 9. Accuracy curve of ALBERT.

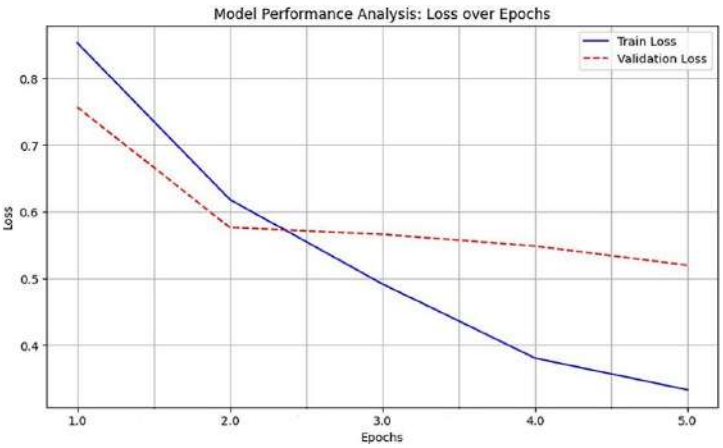

Figure 8. Loss plot for RoBERTa.

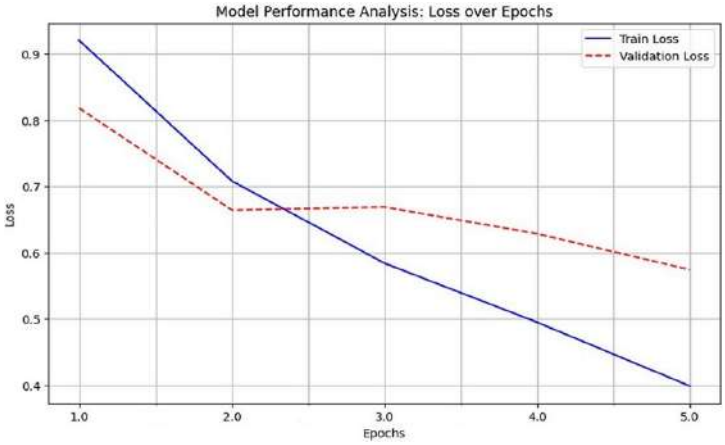

Figure 10. ALBERT loss convergence.

of the RAG models.

The performance of DeepSeek R1 in this context is consistent with comparative analyses such as BytePlus [40], highlights the advantage of DeepSeek R1 compared to other RAG setups in certain cases. Chitika [41] explores its SLM-based optimization with trade-offs to its factual precision. Elephas [42] provides a direct comparison between DeepSeek R1 and Mistral, concluding that Mistral shows better flexibility in zero-shot tasks, albeit at reduced classification accuracy.

4.3.1 Explainability with LIME (RAG Models)

LIME was also employed to explore transparency in RAG predictions:

LLaMA3 (Figure 17): classified the example as a true news with 95% confidence, but words like “Eiffel” “explosion”, “police”, “suspect”, “terror” and “attack” were highly weighted.

DeepSeek R1 (Figure 18): Correctly identified a false rumor with 80% of confidence using words such as “The”, “been”, “set”, “after” and “attack.”

Mistral (Figure 19): also classified the example as a true news with 90% confidence. Overall, fine-tuned transformer classifiers in this experimental environment performed better in the detection score compared to the zero-shot RAG-based LLM baselines. The incorporation of auxiliary tweet metadata improved the model’s contextual understanding and classification accuracy, as it was also the case in previous neural misinformation research [38, 39, 43]. Although RAG models provide on-line explainability and source recovery, the performance in instantaneous, noisy text classification is still low when compared to specialized transformer architectures [44].

Table 4. Detection Results (%) Using RAG-Based Models

Sr. No	Model	Accuracy	Precision	Recall	F1 Score
1	DeepSeek R1	35.00	26.00	35.00	29.00
2	LLaMA3	30.00	22.00	30.00	25.00
3	Mistral	25.00	20.00	25.00	22.00

Figure 11. Performance of Transformer-Based Models for Rumor Detection

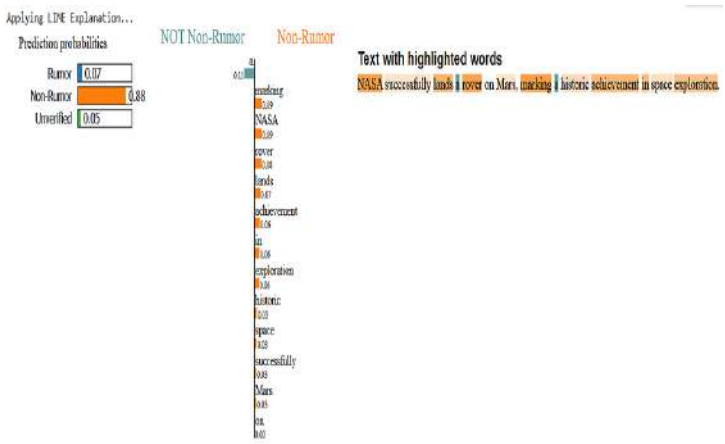

Figure 12. BERT LIME explainability example.

4.4 Experimental Comparisons of all Models

The comparative evaluation of the use of transformers and retrieval-augmented generation (RAG)-based Large Language Models was conducted under different training paradigms. The transformer models were trained on labeled PHEME data, but the LLMs were tested in a zero-shot context on a RAG pipeline without being trained on the labels of the dataset. In this experimental design, fine-tuned transformer models scored better in classification than zero-shot RAG-based LLMs, thus, the performance difference should be regarded as a result of the training design but not structural excellence. Transformer models performed the best under this experimental setting by achieving the highest accuracy (82.95%), followed by RoBERTa (81.50%), DistilBERT, and ALBERT (80.67%). RAG models had lower performance, with DeepSeek R1 (35%), LLaMA3 (30%), and Mistral 25% accuracy. Transformer models showed stable learning behavior, while RAG models had unstable learning curves and higher variance. Explainability analyses using LIME revealed that transformer models pay attention to semantically meaningful words, while RAG models often emphasized irrelevant tokens, making them less reliable in sensitive use cases.

The comparatively lower performance of RAG-based Large Language Models in the considered experimental setting in comparison to the fine-tuned transformer classifiers can be attributed to a number of task-specific reasons. First, rumor detection in social media depend on brief, noisy, and event-driven text, in which significant cues may be subtle and local in context. Retrieval-based augmentation can also add in such environments irrelevant or loosely correlated information, which can be used to dilute rather than improve classification accuracy. Second, Large Language Models are mainly trained to perform generative reasoning problems, and rumor detection is discriminative with multiple-classes classification problems that demand fixed probability estimation. Such an imbalance may cause inaccurate predictions and low classification accuracy. Thirdly, RAG pipelines provide an extra variability in the form of retrieval quality, prompt construction, and probabilistic

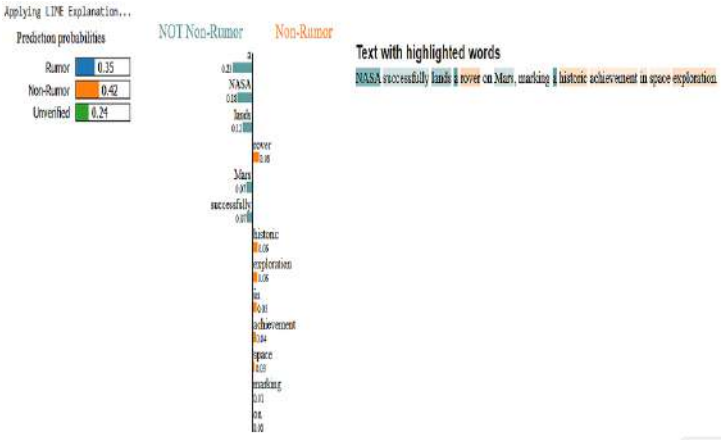

Figure 13. DistilBERT LIME example.

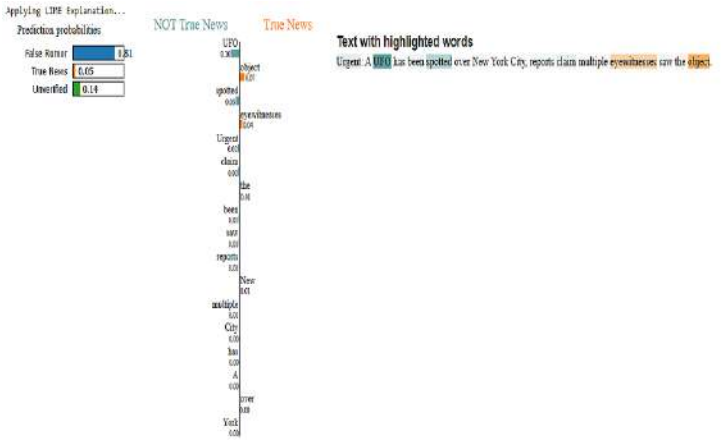

Figure 14. RoBERTa LIME explanation.

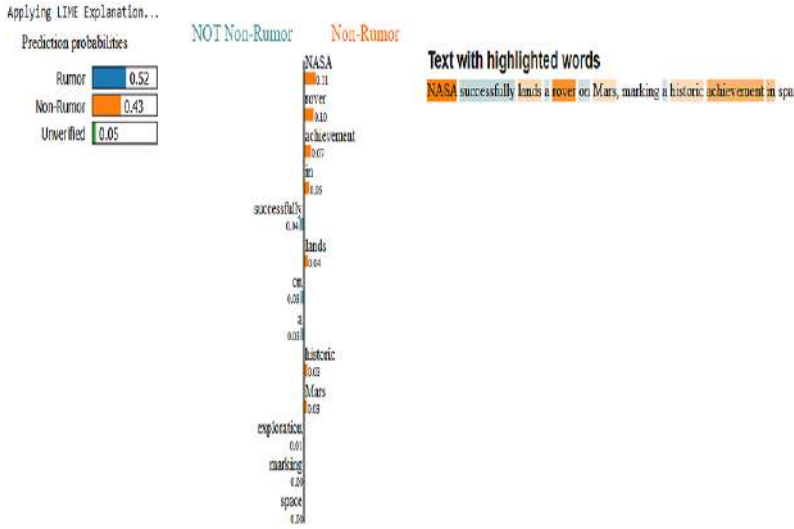

Figure 15. ALBERT LIME visualization.

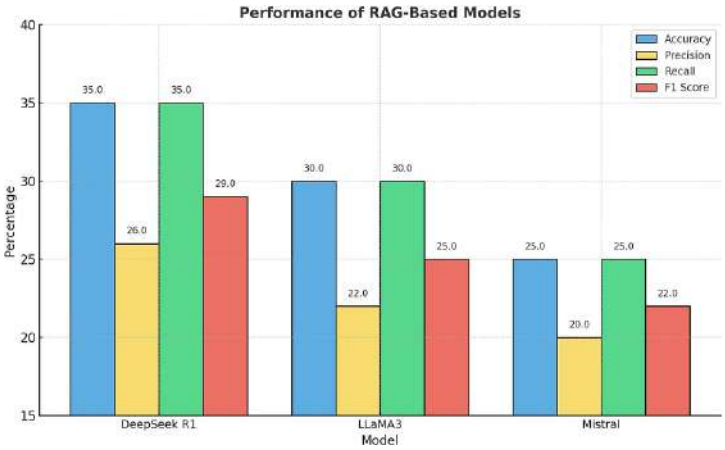

Figure 16. Performance of RAG-Based Models for Rumor Detection

generation which adversely affects reliability and reproducibility. These results are aligned with the previous research that stated that RAG-based methods have been effective in knowledge-rich or long context tasks but do not achieve as high performance in short text classification scenarios, especially in social media usage at real-time. As a result, although the use of RAG-based models can be beneficial to contextual inference and on-line explainability, fined tuned transformer based architectures are still better suited to perform reliable rumor detection tasks, which require accuracy, stability, and explanation.

4.5 Limitations

A few methodological considerations are relevant when interpreting the results of this study. First, the experimental comparison considers the fine-tuned transformer-based classifiers and RAG-based Large Language Models that are tested not based on supervised training on the specific tasks, thus, the observed differences in the performances should be interpreted in the framework of the corresponding training paradigms and not as direct architectural comparisons. Second, the evaluation of RAG-based LLMs was conducted on a relatively small test set ($n \approx 48$), which can impact the statistical strength and external validity of the results. Lastly, retrieval quality, prompt formulation, and generative variability also have the potential to affect the performance of RAG-based models, which presents other factors that should be considered in comparison to supervised fine-tuned classifiers.

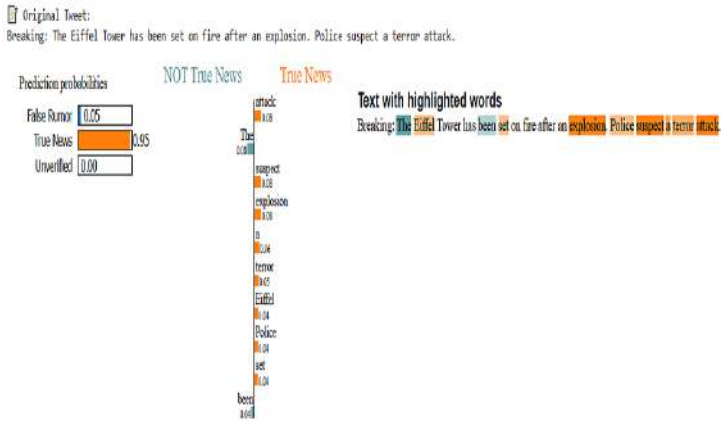

Figure 17. LLaMA3 LIME analysis.

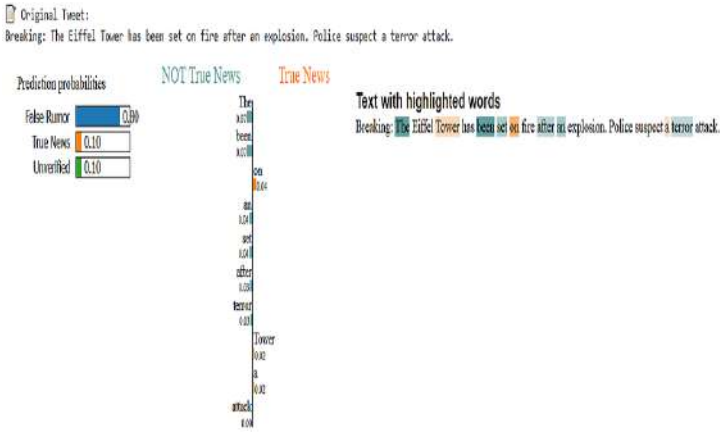

Figure 18. DeepSeek R1 explainability.

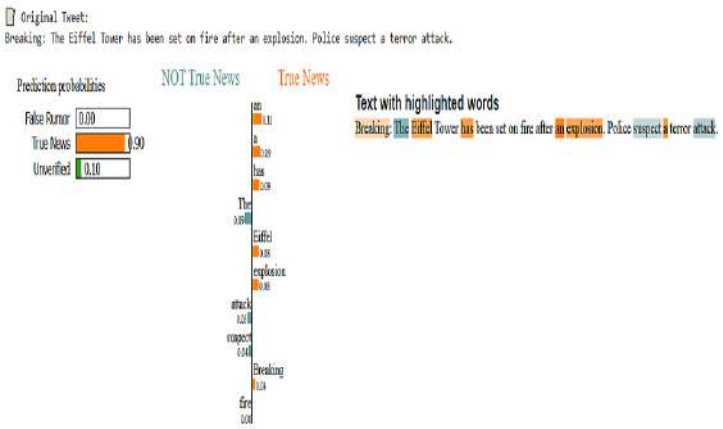

Figure 19. Mistral interpretability chart.

5 Conclusion

This study set out to evaluate and compare the effectiveness of transformer-based models (BERT, RoBERTa, DistilBERT, ALBERT) and retrieval-augmented generation (RAG)-based models (LLaMA3, DeepSeek R1, Mistral) in detecting rumors in social platforms, using PHEME dataset. As we have used both conventional performance metrics such as accuracy, precision, and recall, F1-score as well as explainability techniques such as LIME to uncover the mechanisms of predictions of the models, by this we have not only measured the classification performance of these models but also examined how and why they make predictions.

Fine-tuned transformer based models outperformed zero-shot RAG-based LLMs in the considered experimental environment. Specifically, the highest accuracy of BERT (82.95) was shown to be noted as supervised fine-tuning, which indicates the usefulness of task-specific training in rumor detection. RoBERTa, DistilBERT and ALBERT also showed competitive results at the cost of interpretation-preference and efficiency. The application of LIME established the fact that these models were always able to pay attention to contextually relevant words in the input, thus improving their trustworthiness. In contrast, RAG-based models such as DeepSeek R1, LLaMA3, and Mistral provided promising abilities in real-time information retrieval and long-context reasoning but had difficulties dealing with the short and noisy format of social media text. Their lower performance in this study can be explained by the presence of the zero-shot setting of evaluation and by the generative nature of LLMs, which is not tasked to be increased to the short and discriminative multi-class classification tasks.

One of the key parts of this study is explainable AI (XAI), especially LIME. It allowed for transparent model auditing, which showed not only what the models forecasted but also how they got to their decisions. This is particularly crucial in detecting misinformation, for which accountability and interpretability are as valuable as the accuracy. Moving forward, a few possibilities can be considered for taking the further work. Future work could take the form of the development of hybrid architectures that would bring together the best of transformer and RAG models, which would incorporate robust classification with knowledge-grounded retrieval enhancing both accuracy and the depth of context.

Finally, considerations in assessing fairness, adversar-

ial robustness, and ethics should be prioritized such that such models should not only aim at accuracy in the real world scenarios but also equity and trustworthiness.

Author Contributions

Aziz Ahmad: Conceptualization, Methodology, Software, Writing – Original Draft. **Shams Ur Rahman**: Supervision. **Spogmay Yousafzai**: Data Curation, Investigation, Visualization. **Ghulam Hafeez**: Software, Validation, Writing – Review & Editing.

Compliance with Ethical Standards

The authors declare that they have no conflict of interest. This study uses the publicly available PHEME dataset, which is a collection of anonymized social media posts that were collected for research purposes. No private or personally identifiable information was accessed or processed. This research does not involve direct interaction with human participants and animals. While social media data can be prone to demographic, linguistic or topical biases, explainable AI methods such as LIME are used to ensure transparency and facilitate the responsible interpretation of model outputs.

Funding Information

No external funding was received for this study.

References

[1] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PLoS ONE*, vol. 11, no. 3, p. e0150989, 2016.

[2] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences*, vol. 497, pp. 38–55, 2019.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[4] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] Z. Lan *et al.*, "ALBERT: A lite BERT for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT," *arXiv preprint arXiv:1910.01108*, 2019.

[7] A. Grattafiori *et al.*, "The LLaMA 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[8] D. Guo *et al.*, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[9] F. Jiang, "Identifying and mitigating vulnerabilities in LLM-integrated applications," Master's thesis, Univ. Washington, 2024.

[10] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144.

[12] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample-path-based approach," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 408–421, Feb. 2014.

[13] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "K-center: An approach on the multi-source identification of information diffusion," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2616–2626, Dec. 2015.

[14] B. Malhotra and D. K. Vishwakarma, "Classification of propagation path and tweets for rumor detection using graphical convolutional networks and transformer-based encodings," in *Proc. IEEE 6th Int. Conf. Multimedia Big Data (BigMM)*, New Delhi, India, Sept. 2020, pp. 183–190.

[15] C. M. M. Kotteti, X. Dong, and L. Qian, "Ensemble deep learning on time-series representation of tweets for rumor detection in social media," *Applied Sciences*, vol. 10, no. 21, p. 7541, 2020.

[16] Q. Huang, C. Zhou, J. Wu, L. Liu, and B. Wang, "Deep spatial–temporal structure learning for rumor detection on Twitter," *Neural Comput. Appl.*, vol. 35, no. 18, pp. 12995–13005, 2023.

[17] H. K. Thakur, A. Gupta, A. Bhardwaj, and D. Verma, "Rumor detection on Twitter using a supervised machine learning framework," *Int. J. Inf. Retrieval Res. (IJIRR)*, vol. 8, no. 3, pp. 1–13, 2018.

[18] O. Mairaj and S. U. R. Khan, "Unveiling temporal patterns in information for improved rumor detection," *Social Netw. Anal. Mining*, vol. 15, no. 1, p. 13, 2025.

[19] T. Chen, H. Chen, and X. Li, "Rumor detection via recurrent neural networks: A case study on adaptivity with varied data compositions," in *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD)*, Cham, Switzerland: Springer, June 2018, pp. 121–127.

[20] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on Twitter with hybrid CNN and RNN models," in *Proc. 9th Int. Conf. Social Media and Society*, Copenhagen, Denmark, July 2018, pp. 226–230.

[21] K. Zhou, C. Shu, B. Li, and J. H. Lau, "Early rumour detection," in *Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, June 2019, pp. 1614–1623.

[22] J. Ma, W. Gao, and K. F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2017, pp. 708–717.

[23] Z. Wang and Y. Guo, "Rumor events detection enhanced by encoding sentimental information into time series division and word representations," *Neurocomputing*, vol. 397, pp. 224–243, 2020.

[24] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, 2021.

[25] R. Anggrainingsih, G. M. Hassan, and A. Datta, "BERT-based classification system for detecting rumours on Twitter," *arXiv preprint arXiv:2109.02975*, 2021.

[26] S. Sharma and R. Sharma, "Identifying possible rumor spreaders on Twitter: A weak supervised learning approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, July 2021, pp. 1–8.

[27] B. Pattanaik, S. Mandal, R. M. Tripathy, and A. A. Sekh, "Rumor detection using dual embeddings and text-based graph convolutional network," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 86, 2024.

[28] G. Joshi *et al.*, "Explainable misinformation detection across multiple social media platforms," *IEEE Access*, vol. 11, pp. 23634–23646, 2023.

[29] L. M. S. Khoo, H. L. Chieu, Z. Qian, and J. Jiang, "Interpretable rumor detection in microblogs by attending to user interactions," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 5, Apr. 2020, pp. 8783–8790.

[30] A. Radford *et al.*, "Language models are unsupervised multitask learners," OpenAI Blog, vol. 1, no. 8, p. 9, 2019.

[31] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[32] A. Zubiaga, G. Wong Sak Hoi, M. Liakata, and R. Procter, "PHEME dataset of rumours and non-rumours," figshare Dataset, 2016. doi: 10.6084/m9.figshare.4010619.v1.

[33] C. M. M. Kotteti, X. Dong, and L. Qian, "Ensemble deep learning on time-series representation of tweets for rumor detection in social media," *Applied Sciences*, vol. 10, no. 21, p. 7541, 2020.

[34] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proc. 12th ACM Int. Conf. Web Search and Data Mining (WSDM)*, Jan. 2019, pp. 312–320.

[35] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[36] Hugging Face, "sentence-transformers/all-MiniLM-L6-v2," [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[37] A. Bilal, D. Ebert, and B. Lin, "LLMs for explainable AI: A comprehensive survey," *arXiv preprint arXiv:2504.00125*, 2025.

[38] R. Anggrainingsih, G. M. Hassan, and A. Datta, "Evaluating BERT-based language models for detecting misinformation," *Neural Comput. Appl.*, vol. 37, no. 16, pp. 9937–9968, 2025.

[39] R. Anggrainingsih, G. M. Hassan, and A. Datta, "Evaluating BERT-based pre-training language models for detecting misinformation," *arXiv preprint arXiv:2203.07731*, 2022.

[40] BytePlus, "DeepSeek-R1 vs LLaMA 3 for RAG: A detailed comparison," 2025.

[41] Chitika, "Can DeepSeek's SLM approach replace traditional RAG?" 2025.

[42] Elephas, "Mistral 7B vs DeepSeek R1 performance: Which LLM is the better choice?" 2025.

[43] A. Khraisat, M. Chang, L. Chang, and J. Abawajy, "Survey on deep learning for misinformation detection: Adapting to recent events, multilingual challenges, and future visions," *Social Science Computer Review*, 2025.

[44] Deakin University Digital Repository, "Cross-domain performance of RAG models in fact-checking and social media analysis," 2025.