



RESEARCH ARTICLE

OPEN ACCESS

LEVERAGING LARGE LANGUAGE MODELS FOR RUMOURS DETECTION IN SOCIAL MEDIA

Priyanshi Borase^{*1} and Satish Kolhe²

^{1,2}School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, India.

¹ <http://orcid.org/0009-0009-1282-8239> , ² <http://orcid.org/0000-0002-8423-9152> 

E-mail: *priyanshiborase@gmail.com, srkolhe2000@gmail.com

ARTICLE INFO

Article History

Received: February 04, 2026

Reviewed: March 04, 2026

Accepted: April 10, 2026

Published: June 30, 2026

Keywords:

Rumor Detection,
Social Media,
Natural Language Processing,
Large Language Models,
GPT-4.

ABSTRACT

The rapid increase in the content generated by users on the social media has significantly accelerated the dissemination of rumors and misinformation, creating serious societal challenges. This paper introduces a robust and scalable framework for real-time rumor detection powered by Large Language Models, like BERT, RoBERTa, and GPT-4. The proposed system combines Natural Language Processing techniques with sentiment analysis, stance detection, and automated fact-checking to enhance contextual understanding and assess credibility more effectively. Data is collected from various social media platforms like Twitter, Facebook, and Reddit, along with benchmark datasets like PHEME and FakeNewsNet. Experimental results show that LLMs significantly outperform traditional machine learning models, with GPT-4 achieving an accuracy of up to 94.7%. An ablation study further highlights the contribution of each module to the overall performance of GPT-4 for Rumor Detection. This framework offers a very comprehensive and flexible approach for countering misinformation across a range of languages and social environments.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The popularity of social media has led to massive content production by users. Although these platforms provide the exchange of real-time information, they are also good platforms where misinformation and rumours propagate because there is no centralised control of content. This is a big danger, particularly in times of crisis, e.g. pandemic, election, and natural disasters, when misinformation may cause the population to panic, polarize or end up being damaged. A conventional Natural Language Processing (NLP) system is not entirely applicable in the dynamic web space. They are based on handcrafted superficial lexical properties and do not support the fine-grained context, sarcasm, language changes, and cross-domain linguistic variability. Additionally, feature engineering is not scalable and consumes time when it comes to exposure to the fluid and diverse nature of social media information. Conversely, recent advances in Large Language Models (LLM) such as Generative Pretrained Transformer 4 (GPT-4) and Bidirectional Encoder Representations of Transformers (BERT) make RoBERTa capable of tracking the nuances of human discourse by effectively processing linguistic context, recognising semantic relationships. They are the best candidates for complex problems that require generalisation in tasks, as they are able to learn with massive corpora and combine them to solve other problems. This research paper presents a strong, end-to-end architecture that uses LLMs, as well as support systems, such as stance detection, sentiment analysis, and automated fact-checking, to enhance the especially false or misleading content detection. This paper goes beyond the classification and proposes a modular and real-time rumour detector with fine-tuned LLMs like BERT, RoBERTa, GPT-4, stance detection using conversation thread analysis, sentiment profiling, claim extraction combined with fact-checking APIs, and temporal propagation analysis to detect rumours early. This work distinguishes itself by transforming Large Language Models into a fully integrated end-to-end rumour detection pipeline. The framework combines a variety of complementary signals, rather than only using textual classification; semantic with the help of LLM-based classification, behavioural with the help of stance analysis in the thread of conversation, emotional with the help of sentiment analysis, external fact validation with the help of automated fact-checking sources, and temporal propagation dynamics that investigate information diffusion over time.

These dimensions together allow the system to provide a more context-sensitive and trustworthy rumour detection system than traditional systems based on a transformer. The main aim is to establish a stable and scalable approach that is capable of detecting and mitigating the effect of misinformation as it occurs. The approach will leverage both a profoundly contextual language model and real-time data pipelines to enable a high level of efficiency over a variety of social media and also a variety of languages [1]. The models that are used in the work include BERT, RoBERTa, and GPT-4, which are not innovative, but the implementation that coordinates all of them into a single, real-time rumour detection pipeline is. In contrast to previous research that measures transformer models as independent classifiers, our design incorporates a hybrid multi-stage architecture to combine five complementary dimensions of rumour evidence, namely semantic interpretation with the help of LLM-based classification, behavioural, i.e. stance analysis-based, cues, emotional, i.e. sentiment analysis-based, intensity, and temporal, i.e. temporal propagation-based, models. Such an integrated architecture allows detecting rumours dynamically during the diffusion cycle and not just after the hoc classification. The new pipeline design, cross-module fusion strategy, and the ability to deploy in real-time are novel, as opposed to coming up with a new language model.

II. REVIEW OF LITERATURE

Social media rumour detection is a fast-developing branch of NLP, which is because online content is dynamic, informal and highly context-dependent. Initial research used rule-based methods and classical machine learning models based on manually engineered features derived from message content, user metadata and network structure. Random Forests, Logistic Regression, Support Vector Machine (SVM), and Decision Trees were used to classify suspicious messages using features like length of the message, use of punctuations, sentiment polarity, and user credibility scores. Classifiers based on credibility have been designed by Castillo et al. to break news events [2] and by Kwon et al. to break the real-time credibility of tweets [3]. Even though feature-engineered models were interpretable, they were limited in a number of ways. To begin with, they could not represent sequential and contextual dynamics, and thus could not be used to represent changing narratives or conversational flow [4]. Second, they were specific and intensive to resources, making them have a narrow generalisation across platforms and datasets. They could not also get at more semantic nuances, like sarcasm, idiomatic expressions, or long-range dependencies, due to their shallow representations. Subsequent research took up the Latent Dirichlet Allocation (LDA) unsupervised topic modelling methods, as well as vector-related models of topic modelling, including the Term Frequency Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe. These methods provided better semantic concern as compared to classical feature engineering.

Yet, they continued to fail to deliver dynamic context and word-order embedding's such as Word2Vec and GloVe, which attribute fixed vectors to words without paying attention to context and result in ambiguous signifying polysemous words and discourse nuances [5]. The emergence of deep learning, including such inventions as recurrent neural networks (RNNs), long-short-term memory (LSTM), or convolutional neural networks (CNNs), represents a turning point in the study of rumour detection. All these architectures allowed models to learn features on raw text automatically and comprehend more of the temporal dependence with contents created by users [6]. Such architectures allowed one to automatically learn features of text and enhanced the ability to learn temporal relationships in user-created content. However, long-range dependencies and parallel computation constraints are still problematic with these models. Transformer-based models have been used to revolutionise the NLP community, including BERT [7], RoBERTa [8], and GPT. The models with this initial training on large unlabelled datasets and then used as task-specifically fine-tuned significantly enhanced context modelling, bidirectional dependency understanding and cross-domain generalisation. An example was the deep bidirectional, self-attention architecture of BERT, which was capable of performing better than previous models in various tasks, including GLUE and SQuAD, due to its contextual representations.

Similarly, GPT and its successors, particularly GPT -3, were shown to have strong few-shot and zero-shot learning performance, including performance equivalent to other NLP tasks using a small number of task-specific training examples [9]. More recent literature has suggested the incorporation of LLMs into modular models that combine stance detection, sentiment analysis, and fact-checking to come up with more precise rumour classification. To effectively represent the target-specific representations and eliminate target-invariant noise, Sun et al. suggest a multi-target adversarial attention network approach to stance detection, which achieves better results on a variety of stance detection benchmarks. [10]. On the same note, Zhou et al. pointed out that by incorporating sources of fact-checking and knowledge, such as Snopes, PolitiFact, or Wikipedia, which enhance the model to measure the consistency of facts [11]. The supporting features that have been investigated include emotion detection and discourse analysis to improve performance. Multitask learning schemes have been constructed so as to learn rumour detection, stance classification, and source reliability estimation simultaneously. These methods assist in improved generalization of datasets and platforms. Ma et al. suggested the use of a tree-based recursive neural network in order to predict the propagation patterns of rumours in Twitter threads [12]. Nevertheless, although they have good performance, there are no challenges to LLM-based approaches.

Its expensive computational nature, the potential to over fit on biased data, and vulnerability to adversarial examples are all serious issues [13]. Furthermore, explain ability and transparency of deep models remain pressing issues, especially in high-stakes applications such as misinformation mitigation during elections or health crises. Recent work has moved towards the multimodal approach of misinformation detection, where image-text or audio-text pairs are considered in a joint manner instead of using textual cues only. Zeng et al. [14] proved that synthetic multimodal training examples are an important way to enhance generalisation in the real-life misinformation context. Rumour detection based on vision-language is still developing, and Tahmasebi [15] demonstrates that large vision-language models somehow outperform all-textual rumour detection systems. On the same note, Shen et al. [16] used contrastive learning with optimal transport to enhance the correspondence between visual and verbal features, whereas Zhu et al. [17] created a novel multimodal dataset that facilitates the detection of misinformation as well as the localisation of deceptive content in the image. Liu et al. [18] went ahead to broaden the field to include the audio cues in detecting manipulated multimedia misinformation. Our study differs by operationalising Large Language Models into a complete end-to-end rumour detection pipeline that extends beyond text classification.

In addition to semantic interpretation through LLM-based classification, the system integrates behavioural cues derived from stance analysis within threaded conversations, emotional patterns identified through sentiment profiling, and external factual verification using automated fact-checking sources. Furthermore, the framework analyses temporal propagation patterns to capture how information evolves and spreads over time. By combining semantic, social, emotional, factual, and temporal intelligence, the system enables more reliable and context-aware rumour detection compared to traditional transformer-only approaches. In summary, the literature demonstrates a clear trajectory from feature-based models to sophisticated transformer-based LLMs, with increasing emphasis on context modelling, knowledge integration, and real-time processing. The fusion of LLMs with auxiliary NLP modules represents the latest advancements in rumour detection, offering significant gains in performance and robustness over traditional approaches.

III. METHODOLOGY

This section provides a systematic approach towards reaching the research objectives. It describes the data collection and the Model Architecture. The methodology combines a quantitative analysis with the sophisticated methods of the LLM and aim to preserve the reliability, scalability and reproducibility of the proposed rumour detection framework.

III.1 DATA COLLECTION

Social media materials were sourced on popular sites through Twitter, Facebook and Reddit to have strong and varied training data to detect rumours. The reason behind their choice is that these platforms are widely used, content is disseminated live, and the amount of text created by users is enormous, making it the perfect soil to spread false information. Besides real-time scraping of social media, several publicly available benchmark datasets, as listed below, were used to supplement the training and evaluation processes:

- Covid19_RumourDataset: This dataset is curated throughout the COVID-19 pandemic and includes labelled rumours and factual information that has been verified and is associated with health, vaccine efficacy, lockdowns, and other policy announcements [19].
- PHEME Dataset: Twitter conversation threads that are annotated by veracity and stance of rumour, containing the events of the Ferguson, German wings Crash and Charlie Hebdo attack [20].
- FakeNewsNet: A unified data set that includes the content of the news reports, as well as social dynamics and spatial-temporal features that facilitate misinformation detection studies [21].
- Twitter15 and Twitter16: Labelled Twitter dataset Benchmark data, labelled with different rumours, and reflecting different propagation patterns [12].

The raw data underwent a strict pre-processing step to enhance the quality of the data and remove noise. This involved eliminating non-textual materials (e.g. pictures, links, emojis), spam, duplicates, and adverts. The other pre-processing tasks included language identification, tokenisation, lemmatisation, and normalisation to maintain the textual consistency between different sources of data. The obtained cleaned corpus was used to fully train and test the suggested rumour detection framework.

III.2 MODEL ARCHITECTURE

The suggested rumour detection model takes a hierarchical and modular structure that is intended to process high volumes of noisy and context-sensitive data from social networking sites. Figure 1 shows the model architecture of the proposed system.

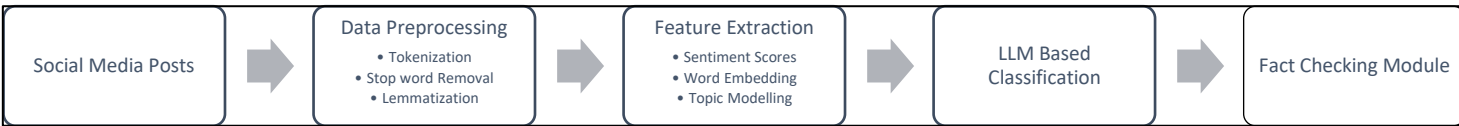


Figure 1: Model Architecture of Proposed System.
Source: Authors, (2026).

III.3 DATA PROCESSING

Before textual data is introduced into downstream machine learning models, a rigorous pre-processing pipeline is applied to normalise and standardise the content. It starts with tokenisation, which separates raw text into individual words or tokens with highly established NLP tools like SpaCy and the Natural Language Toolkit (NLTK) [22,23]. Then, stopwords removal is carried out to remove common function words (e.g., the, is, at, etc.) that usually have little semantic content [24]. This is followed by lemmatization which involves converting the words into their base or dictionary form, such as changing running into run. The action increases the consistency of lexes and minimises the vocabulary [25]. All of these pre-processing methods serve to eliminate the linguistic noise and only preserve the most relevant semantic meaning, thus allowing more effective feature extraction and overall better performance of the model as it trains.

III.4 FEATURE EXTRACTION

Various sophisticated feature extraction methods are used to extract more linguistic, semantic and topical features of the information. Sentiment scores are calculated to measure the tone and bias of emotion in the text, in tools including VADER [26] and TextBlob [27], to categorize a piece of writing as either positive, negative, or neutral. Moreover, there are word embedding's that are produced with models such as Word2Vec [28] and FastText [29], and create dense vector representations of words based on their non-relatedness in context. The embedding allows the model to make reasonable semantic similarity and a cross-synonymic and cross-contextually relevant generalization.

Moreover, LDA, a generative probabilistic model, is used to carry out topic modelling and thus identify latent thematic structures within the corpus [30]. This would help in identifying topic-specific rumors and also increase the capability of the model in differentiating the various kinds of misinformation. Together, the characteristics give rich, semantically meaningful inputs to the main classification module and assist auxiliary implementations like stance detection.

III.5 LLM-BASED CLASSIFICATION

The main element of the offered framework applies the use of LLMs to the high-accuracy classification of the social media posts as rumors or non-rumors. Variations of a few of the most recent developments in the transformer-based modelling are fine-tuned in order to encode the salient linguistic and contextual hints of social discourse. BERT can be applied in its bidirectional attention mechanism, which allows for more comprehension of the bidirectional context in a sentence. BERT can be optimized to yield a stronger, highly optimized model called RoBERTa, which outperforms BERT by using dynamic masking and training models on multiple large corpora. It is also equipped with GPT 4 due to its generative ability and remarkable performance in few-shot and zero-shot learning, which presents high versatility in understanding various domains of languages. Labelled datasets are used to fine-tune these models and optimise the cross-entropy loss. Standard measures are used to evaluate the performance, such as accuracy, precision, recall, and F1-score, to be able to obtain reliable and generalizable classification results.

The collective position of the social discourse is very often used as the basis of rumor verification. In this module, conversation threads or replies are analyzed to find the position as compared to the original post. The classification includes:

- Support: The reply agrees with or reinforces the original post.
- Deny: The reply refutes or contradicts the claim.
- Query: The reply asks for clarification or raises questions.
- Comment: The reply provides commentary without explicit support or denial.

Knowledge of stance assists in combining the opinion of the community and enhances classification strength, especially in situations when the initial assertion is unclear. A fact-checking module is incorporated to ensure that claims are factual. This component is the one that finds the main claim of the input text with the help of claim detection algorithms. Makes cross-references to reputable external databases like Snopes, PolitiFact and Wikipedia, applying similarity matching techniques and entailment techniques. Output is a credibility score or flag that shows that the assertion is confirmed, disputed, or insufficiently informed. The fact-checking feature introduces another level of verification, which makes the system more credible and resistant to new and unfamiliar claims.

III.6 ANALYSIS

Temporal Analysis is seizing Evolutionary Dynamics of Misinformation Propagation with Time. In order to be more aware and identify rumors in their flow, this framework adds a specialized temporal analysis component, which simulates the development of misinformation with time and network topology. Temporal modelling is better than predictive (to detect it in time) and better than interpretable. This is primarily aimed at modelling time-varying propagation patterns of individual claims (including cascades and threads), derivation of temporal representation of velocity, acceleration and half-life to augment content-based LLM signals, and temporal modelling of early warnings and assessment of intervention strategies. In ingesting data, timestamps and conversation format, such as original posts, replies, and shares, are maintained in order to form ordered cascades so that every post can be represented as user, time, and attributes. Public datasets such as PHEME, FakeNewsNet, and Twitter15/16 provide the required temporal fields.

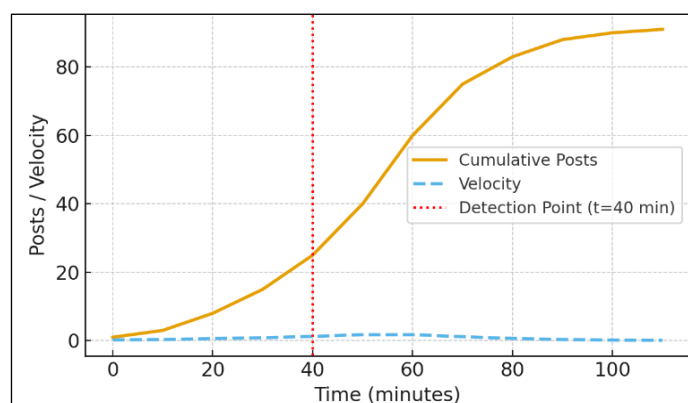


Figure 2: Temporal Analysis of Misinformation Propagation
Source: Authors, (2026).

Temporal feature engineering is a feature engineering that is aimed at creating volume curves, cumulative posts, velocity and detection point like in Figure 2. The approaches to temporal modelling are nonparametric analysis and descriptive (e.g., kernel smoothing and change-point detection), Hawkes processes of self-exciting propagation, dynamic graph embedding and temporal graph networks, sequence models (e.g. LSTM or TCN with time-aware LLMs), and cascading diffusion classifiers (e.g. early detection). Evaluation metrics include time-to-detection, temporal AUC, early-detection F1@T (e.g., at 10, 30, or 60 minutes), and intervention effect simulation. Altogether, the temporal analysis will be beneficial in complementing the semantic understanding of LLM by allowing to detect misinformation in advance, revealing diffusion, and determining the most favourable time to intervene.

IV. EXPERIMENTAL SETUP AND IMPLEMENTATION

The proposed rumour detection system is implemented in Python programming language, with the use of key libraries, such as scikit learn to use traditional machine learning techniques and PyTorch to implement and fine-tune deep learning techniques. Data processing and feature engineering tasks are facilitated by NLP libraries including SpaCy, NLTK, and HuggingFace Transformers.

IV.1 TRAINING ENVIRONMENT AND INFRASTRUCTURE

Training and evaluation of the model are done through the assistance of labelled datasets through Twitter and other social media sites, along with benchmark datasets like PHEME, Covid19_RumourDataset and Twitter15/Twitter16. These datasets have manually annotated labels, which are used to identify posts as rumours or non-rumours, which are used to perform supervised learning. All datasets were divided with an 80-10-10 stratified split to obtain training, validation and test sets to guarantee reproducibility and minimise bias. Stratification was done so that the ratio of rumour to non-rumour labels was the same in all splits as shown in Table 1.

Table 1: Dataset Splits for Datasets.

Dataset	Total Samples	Rumour (%)	Non-Rumour (%)	Used For
PHEME	6,425	45.6%	54.4%	Training / Testing
FakeNewsNet	23,196	51.2%	48.8%	Training
Covid19 Rumour Dataset	10,280	60.3%	39.7%	Validation / Testing
Twitter15 & Twitter16	2,608	53.9%	46.1%	Testing

Source: Authors, (2026).

Hyperparameters are optimized using a combination of grid search (for BERT, RoBERTa) and Bayesian optimization for GPT-4 fine-tuning via parameter-efficient fine-tuning as shown in Table 2.

Table 2. Hyperparameter Tuning.

Parameter	Search Range	Best Value (BERT / RoBERTa)	Best Value (GPT-4 PEFT)
Learning Rate	1e-3 → 1e-6	2e-5	1e-5
Batch Size	8, 16, 32	16	8
Max Sequence Length	128, 256, 512	256	512
Dropout Rate	0.1, 0.2, 0.3	0.1	0.1
Warmup Steps	0, 500, 1000	500	1000

Source: Authors, (2026).

The training is implemented in cloud-based infrastructure with GPU-enabled instances (such as NVIDIA A100 on AWS EC2 or Google Cloud TPU VMs) so that it is possible to train large transformer models in parallel. Scalable, real-time inference with the support of Docker containers and REST APIs to integrate into social media monitoring systems is also supported by cloud deployment. The grid search and random search were used to implement the process of hyperparameter optimization, and parameters such as learning rates, batch sizes and dropout rates were targeted. Early stopping conditions to prevent overfitting are applied. All transformer models were optimized with AdamW and cross-entropy loss. In order to avoid overfitting, an early stopping mechanism with a patience of three epochs was used. The training was carried out depending on model convergence, between 3 and 5 epochs. The learning rate was gradually ramped with a linear learning rate schedule with a warm-up followed by decay. The weight decay was adjusted to 0.01 to normalize the model and reduce overfitting. To prevent gradient explosion during backpropagation, they used gradient clipping with a threshold of 1.0. Mixed-precision training (FP16) was applied during the entire process of fine-tuning to increase the efficiency of training and decrease the computational load.

IV.2 EVALUATION METRICS

Standard classification measures are used to evaluate the performance of each model, to make sure that the evaluation is comprehensive:

- **Accuracy:** This is a measure that determines the ratio of correct instances to the number of instances.
- **Precision:** the proportionality of the number of true positives to the total number of positive predictions, which is a measure of the exactness of the model.
- **Recall:** It is the ratio of true positive predictions of all the actual positive cases, which represent the capability of the model to represent the relevant cases.
- **F1-Score:** It is the harmonic mean of precision and recall, which has a balanced measure, especially when dealing with imbalanced data.

These measures give a general picture of the accuracy as well as the strength of the detection structure. In order to guarantee the credibility of comparisons of model performances, statistical significance analysis was done on the evaluation parameters such as the Accuracy, Precision, Recall and F1-score.

All the models were trained and tested with the help of k-fold cross-validation (usually, $k = 5$ or 10), and the performance measures were computed on average over folds. The mean $\pm$ standard deviation has been given to show variability on each metric, whereby the standard deviation is used as an error bar on comparative plots. Paired statistical tests, including the paired t-test or the Wilcoxon signed-rank test, were used to determine whether or not observed differences between models were statistically significant (usually at $p < 0.05$). In case of a multiple model simultaneous comparison, a one-way ANOVA followed by a post-hoc Tukey HSD was used to regulate multiple comparisons. Besides this, bootstrap resampling ($n = 1,000$) was used to compute a 95% confidence interval of mean scores to provide strong estimates of uncertainty. The procedures are used to ensure that the performance improvement is not attributed to chance variation and give a statistically valid ground to compare the models.

V. MODEL TRAINING AND EVALUATION

The following section includes the Model Training and Evaluation parameters. It describes Base line Model, LLM Performance and Ablation study.

V.1 BASELINE MODELS

Naive Bayes and the SVM are traditional NLP-based classifiers that are used as performance baselines. These models are trained based on feature vectors obtained based on TF-IDF and Bag of Words representation. They are computationally efficient, but do not scale to the capacity to capture both contextual and semantic nuances of social media text, which tends to cause worse performance in practice

V.2 LLM PERFORMANCE

The performance of fine-tuned models such as BERT, RoBERTa, and GPT-4 is by far better than the base models. This is because of their capability to model both context, attention mechanisms and deep semantic relationships that result in increased classification accuracy and greater generalisation across datasets. GPT-4 more so demonstrates strong few-shot learning capabilities and the ability to accept a range of language usage, making it an ideal system in rumour identification of evolving rumours. According to the experiment, BERT has a high accuracy of above 90%, which is much higher than the conventional approach. RoBERTa also improves the performance by optimising the pre-training strategies. More scores are detected in GPT-4, especially on the F1-score, due to its huge pre-training on different datasets.

V.3 ABLATION STUDY

An ablation study is conducted to understand the contribution of individual components in the proposed architecture. The evaluation of the LLM-based classification pipeline is done by selectively disabling the stance detection, sentiment analysis and fact-checking modules. It eliminates the results of fact-checking by a 5-7% loss of accuracy, which makes it significant in proving assertions. The state of disabling stance detection interferes with the process of community response interpretation by the model, especially in threaded dialogue. Lack of sentiment analysis makes the accidental misclassification of the emotionally colored, but true content, less accurate. The obtained ablation results prove that each of the modules plays a very significant role in the overall performance of the rumour detection framework.

VI. RESULTS AND DISCUSSIONS

The experimental findings distinctively reveal that the performance of the LLMs is considerably better compared to the traditional NLP classifiers in rumour detection activities. Table III. is a comparative outcome of model performance on the evaluation dataset. GPT-4, which was fine-tuned on rumour detection as one of the tested models, had the best overall metrics with an accuracy of 94.7% and an F1-score of 94.1% as in Table 3. These findings indicate that GPT-4 has high levels of generalisation and has a better grasp of linguistic and situational contexts. The classical models, such as Naive Bayes and SVM, have performed poorly to moderate levels with an F1-score of 70.1% and 74.6%, respectively. Such models could not capture the subtle linguistic patterns and dependencies, and were particularly ineffective in informal or sarcastic text. Models of the fine-tuned transformer, such as BERT and RoBERTa, showed outstanding performance, with RoBERTa a little ahead of BERT because of the optimized pre-training process. BERT had an F1-score of 88.4% with a 90.5% score of RoBERTa. An ablation study was conducted on the entire pipeline to determine the importance of each part. Performance by the fact-checking component was reduced by about 6.3% in terms of F1-score when the fact-checking component was removed. The removal of stance detection and sentiment analysis reduced by 4.2% and 3.5%, respectively. These findings justify the need to add auxiliary modules in order to improve the model performance. Also, qualitative analysis showed that LLMs are more inclined to notice subtle misinformation and less prone to being misguided by deceptive linguistic patterns or fashionable sentiment tendencies.

Table 3. Evaluation Metrics.

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	71.2%	70.4%	69.8%	70.1%
SVM	75.5%	74.9%	74.3%	74.6%
BERT (Fine-tuned)	90.79%	88.7%	88.1%	88.4%
RoBERTa (Fine-tuned)	91.5%	90.8%	90.3%	90.5%
GPT-4 (Fine-tuned)	94.7%	94.3%	93.9%	94.1%

Source: Authors, (2026).

VII. CONCLUSION

The proposed framework can be highly applicable in the non-experimental environment and is applicable to be deployed in real-time on social media. The system, in contrast to the methods used in the past where emphasis is placed on text classification only, is a combination of stance detection, sentiment analysis, and fact-checking, thus making it possible to identify rumors earlier and within a more context-specific manner. The modular architecture enhances scalability, whereby components can be deployed either in parallel or independently and with ease, by using REST APIs or micro services via Docker. Using multilingual LLMs like GPT-4 and BERT trained with various languages, the framework supports multiple languages without having to invest in training a model, which lowers the cost of annotation and makes the system more flexible in other regions. While inference latency may increase on high-volume data streams and performance may vary for low-resource languages, the results show improved accuracy, efficiency, and multilingual applicability compared to prior models, making the framework suitable for real-world rumor monitoring and decision-support systems.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Priyanshi Borase, Satish Kolhe.

Methodology: Priyanshi Borase, Satish Kolhe.

Investigation: Priyanshi Borase, Satish Kolhe.

Discussion of results: Priyanshi Borase, Satish Kolhe.

Writing – Original Draft: Priyanshi Borase, Satish Kolhe.

Writing – Review and Editing: Priyanshi Borase, Satish Kolhe.

Resources: Priyanshi Borase, Satish Kolhe.

Supervision: Priyanshi Borase, Satish Kolhe.

Approval of the final text: Priyanshi Borase, Satish Kolhe.

VIII. REFERENCES

- [1] T. Sengodan, S. Misra, and M. Murugappan, Eds., *Advances in Electrical and Computer Technologies: Proceedings of the 6th International Conference on Advances in Electrical and Computer Technologies (ICAECT 2024)*, 1st ed. London, U.K.: CRC Press, 2025. doi: 10.1201/9781003515470.
- [2] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th Int. World Wide Web Conf. (WWW 2011)*, Hyderabad, India, Mar. 28–Apr. 1, 2011, pp. 675–684. doi: 10.1145/1963405.1963500.
- [3] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumour propagation in online social media," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dallas, TX, USA, Dec. 7–10, 2013, pp. 1103–1108. doi: 10.1109/ICDM.2013.61.
- [4] R. Anggrainingsih, G. M. Hassan, and A. Datta, "Transformer-based models for combating rumours on microblogging platforms: A review," *Artif. Intell. Rev.*, vol. 57, p. 212, 2024.
- [5] Q. Du et al., "A topic recognition method of news text based on word embedding enhancement," *Comput. Intell. Neurosci.*, vol. 2022, Art. no. 4582480, 2022.
- [6] Y. Xu et al., "Deep recurrent neural network and data filtering for rumor detection on Sina Weibo," *Symmetry*, vol. 11, p. 1408, 2019.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies*, Minneapolis, MN, USA, Jun. 2019, vol. 1, pp. 4171–4186.
- [8] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [9] T. Brown et al., "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.
- [10] Q. Sun, X. Xi, J. Sun, Z. Wang, and H. Xu, "Stance detection with a multi-target adversarial attention network," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, pp. 1–21, 2022.
- [11] X. Zhou, J. Wu, and R. Zafarani, "Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conf. Knowledge Discovery Data Mining*, Singapore, May 6, 2020. Cham, Switzerland: Springer, pp. 354–367.
- [12] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2017.
- [13] R. Bommasani, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.
- [14] F. Zeng, W. Li, W. Gao, and Y. Pang, "Multimodal misinformation detection by learning from synthetic data with multimodal LLMs," *arXiv preprint arXiv:2409.19656*, 2024.
- [15] S. Tahmasebi, E. Müller-Budack, and R. Ewerth, "Multimodal misinformation detection using large vision-language models," in *Proc. 33rd ACM Int. Conf. Information and Knowledge Management (CIKM)*, Boise, ID, USA, Oct. 21, 2024, pp. 2189–2199.
- [16] X. Shen et al., "Multimodal fake news detection with contrastive learning and optimal transport," *Front. Comput. Sci.*, vol. 6, Art. no. 1473457, 2024.
- [17] Y. Zhu, Y. Wang, and Z. Yu, "Multimodal fake news detection: MFND dataset and shallow-deep multitask learning," *arXiv preprint arXiv:2505.06796*, 2025.
- [18] M. Liu et al., "Exploring the role of audio in multimodal misinformation detection," in *Proc. IEEE 14th Int. Symp. Chinese Spoken Language Processing (ISCSLP)*, Singapore, Nov. 7, 2024, pp. 204–208.
- [19] F. Alam et al., "Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and society," in *Findings Assoc. Comput. Linguistics: EMNLP 2021*, Nov. 2021, pp. 611–649.

- [20] A. Zubiaga et al., “Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations,” arXiv preprint arXiv:1609.09028, 2016.
- [21] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media,” *Big Data*, vol. 8, pp. 171–188, 2020.
- [22] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA, USA: O’Reilly Media, 2009.
- [23] M. Honnibal, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,” unpublished work, 2017.
- [24] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., online manuscript. Stanford, CA, USA: Stanford Univ., 2025. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/> (accessed: Aug. 24, 2025).
- [25] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [26] C. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proc. Int. AAAI Conf. Web Social Media*, Ann Arbor, MI, USA, May 2014, vol. 8, pp. 216–225.
- [27] S. Loria, “TextBlob: Simplified text processing,” 2018. [Online]. Available: <https://textblob.readthedocs.io/en/dev/> (accessed: Jan. 15, 2018).
- [28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
- [29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.
- [30] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.