

Explainable Misinformation Detection: A Hybrid Bert And Large Language Model (LLM) Framework for Interpretable Fake News Classification

Deepak Mangal¹, Dr. Gurvinder Singh², Dr. Shashi Shekhar³

¹Research Scholar, Department of CSE, Arni University, Indora, H.P.

²Dean Engineering, Arni University, Indora, H.P.

³Professor, Amity University, Patna, India

ARTICLE INFO

Received: 03 Feb 2024

Revised: 18 March 2024

Accepted: 29 March 2024

ABSTRACT

The rapid growth of online news and social-media platforms has increased the spread of misinformation and created a need for automated detection methods that are both effective and interpretable. This study presents a compact hybrid framework for explainable misinformation detection that combines traditional machine-learning models, Transformer-based classification, post-hoc explainability, and controlled Large Language Model (LLM) interpretation. The LIAR-PLUS dataset is transformed into a binary classification task, with statements categorized as Real or Misinformation. TF-IDF-based Support Vector Machine (SVM) and Logistic Regression (LR), along with DistilBERT and BERT, are evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. BERT is subsequently used as the primary classifier for SHAP and LIME-based local explanations. To examine explanation behaviour and model sensitivity, SHAP-LIME feature overlap, SHAP-based feature deletion, and prediction stability are analysed. The study further employs a controlled local Qwen2.5-0.5B-Instruct model to convert SHAP-derived evidence into human-readable explanations without using the LLM as an independent fact-checker. The experimental results show that traditional and Transformer-based models achieve comparable performance under the evaluated configuration, while SHAP and LIME provide complementary local feature-level interpretations. Feature deletion demonstrates increasing prediction perturbation as more influential features are removed, while most predictions remain stable under the applied perturbation. The controlled LLM can generate structured explanations from model evidence, but unsupported semantic or contextual interpretations may still occur. Therefore, grounding and safety validation remain important when LLMs are used for explainable misinformation detection.

Keywords: Misinformation Detection, Fake News Detection, BERT, SHAP, LIME, Explainable AI, Large Language Models, Qwen2.5, LIAR-PLUS

Introduction

The rapid growth of online news and social-media platforms has made misinformation easier to create, distribute, and amplify. This has increased the need for automated methods that can identify misleading information from textual content. Recent misinformation research has increasingly moved from conventional machine-learning approaches toward contextual Transformer-based models and, more recently, Large Language Models (LLMs). BERT-based models have demonstrated strong contextual representation capabilities for fake-news classification, while recent studies have explored hybrid and enhanced BERT architectures for improving detection performance [1], [2].

However, classification performance alone is insufficient when the system is expected to provide an understandable reason for its prediction. Recent research specifically identifies explainability as an important requirement in fake-news detection and investigates feature-level and reasoning-based explanations [3], [4].

Beyond classification accuracy, practical misinformation-detection systems require explanations that allow users and researchers to understand how a model arrived at an individual prediction. This requirement becomes particularly relevant for Transformer-based classifiers, whose contextual representations are difficult to interpret directly. Post-hoc methods such as SHAP and LIME can expose influential textual features, but different explanation methods may identify different features for the same instance. Moreover, when an LLM is used to convert numerical attributions into natural-language explanations, fluency alone does not guarantee that the generated explanation remains grounded in the original model evidence. Therefore, an integrated evaluation should consider not only classification performance but also explanation behaviour, prediction sensitivity, stability, and the possibility of unsupported LLM-generated interpretations.

The emergence of LLMs has further expanded the possibilities for natural-language interpretation. At the same time, LLM-generated explanations may contain unsupported or hallucinated information, making controlled generation and grounding important for reliable use [5], [6].

Accordingly, this study implements a compact hybrid framework in which BERT performs misinformation classification, SHAP and LIME provide feature-level explanations, and a controlled local LLM converts model-derived evidence into a human-readable explanation.

Problem Statement

Existing misinformation-detection systems primarily focus on improving classification performance. However, a high-performing classifier may still provide limited insight into why a particular prediction was generated. Recent explainable fake-news research highlights this interpretability challenge, while newer LLM-based approaches introduce additional concerns regarding factuality and unsupported generation [3], [4].

The main problems addressed in this implementation are:

Classification without explanation: the predicted class does not directly indicate which textual features influenced the decision.

Limited interpretability of Transformer models: BERT-based classifiers provide contextual representations but their individual predictions are not inherently transparent.

Uncontrolled LLM explanations: LLMs can generate fluent explanations but may introduce information that is not supported by the model evidence [5], [6].

Need for robustness analysis: an explanation should be examined not only for its readability but also for its relationship with model prediction changes.

Therefore, the study focuses on an evidence-grounded and controlled explanation pipeline rather than using an LLM as an unrestricted misinformation classifier.

Objectives

The objectives of this study are:

- To develop a BERT-based binary misinformation classifier.
- To compare traditional machine-learning models with Transformer-based models.
- To generate local explanations using SHAP and LIME.
- To examine explanation consistency and prediction stability through feature-attribution analysis and feature deletion.

To investigate a controlled LLM layer for converting model-derived evidence into concise, human-readable explanations.

State-of-the-art

The main contributions of this study are as follows:

Comparative Classification Framework

A comparative experimental framework is implemented using traditional machine-learning models and Transformer-based models for binary misinformation classification.

SHAP/LIME-Based Local Interpretability

SHAP and LIME are applied to the BERT classifier to identify influential textual features and provide local explanations for individual predictions.

SHAP Feature-Deletion Analysis

A feature-deletion analysis is performed by removing top SHAP-identified features and measuring changes in model probability and prediction stability.

Controlled LLM Explanation Layer

A controlled local LLM layer is implemented to convert SHAP-derived model evidence into concise, human-readable explanations while restricting unsupported interpretations.

Empirical Analysis of Small Local LLM Limitations

The study empirically examines the limitations of small local LLMs in generating model-grounded explanations, including the risk of unsupported or poorly aligned interpretations.

Related Work

Traditional Machine Learning for Misinformation Detection

Traditional misinformation-detection approaches commonly represent textual content using features such as TF-IDF, n-grams, and Bag-of-Words, followed by classifiers such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees, and ensemble methods. Recent studies continue to demonstrate the usefulness of these approaches because of their relatively simple implementation and computational efficiency [1], [3]. A recent machine-learning study, for example, evaluated TF-IDF with multiple classifiers including LR and SVM for fake-news classification [8].

However, traditional feature-based models generally have limited ability to capture deeper contextual and semantic relationships within text. Therefore, they remain useful as baseline models, while Transformer-based approaches provide a stronger representation of contextual information.

Deep Learning and Transformer-Based Detection

Transformer-based language models have significantly influenced misinformation detection by providing contextual representations of textual information. BERT-based approaches have been investigated for fake-news classification, including hybrid architectures that combine BERT representations with conventional machine-learning classifiers [1]. More recent work has also investigated fine-tuning strategies for BERT-based fake-news classification, showing the continuing relevance of Transformer models for this task [2].

Compared with traditional TF-IDF representations, BERT can capture contextual relationships between words and sentences. Smaller Transformer variants such as DistilBERT provide a computationally lighter alternative, making them useful for comparative experimentation. Nevertheless, the internal representations of Transformer classifiers remain difficult to interpret directly, creating a need for post-hoc explainability methods [3].

Explainable AI for Text Classification

Although Transformer models can provide strong predictive representations, their predictions do not automatically provide an understandable explanation. Recent research on explainable fake-news detection emphasizes the importance of identifying the textual evidence that contributes to a model's decision [3]. The literature includes several post-hoc explanation approaches that can identify influential features at the instance level.

In this study, SHAP and LIME are used for local interpretation of BERT predictions. SHAP provides feature-attribution values indicating the contribution of individual features toward a prediction, whereas LIME generates local approximations to explain an individual model decision. Their use enables the proposed system to move from simply reporting “Real” or “Misinformation” toward identifying the textual features associated with that decision.

Recent explainable fake-news research also indicates that explanation quality should not be considered only from the perspective of visual or qualitative interpretation; explanation evaluation and robustness remain important research considerations [3].

LLMs for Explainable NLP

The emergence of Large Language Models has introduced another possibility for generating explanations in natural language. Instead of presenting only numerical feature-attribution values, an LLM can potentially convert model-derived evidence into a concise explanation that is easier for non-technical users to understand.

However, fluent generation does not necessarily guarantee factual or evidence-supported output. Recent research has highlighted hallucination and factuality problems in LLM-generated content [5], [6]. In the misinformation domain, these limitations are particularly important because an explanation that introduces unsupported information can create additional uncertainty rather than improving interpretability.

Recent work such as DELL demonstrates that LLMs can be incorporated into misinformation-detection pipelines for generating reactions, explanations, and task-specific contextual information, while also recognizing the limitations of directly relying on LLMs for news-veracity judgments [7].

Therefore, the present study uses the LLM specifically as a controlled interpretation layer. The LLM receives model-derived SHAP evidence rather than independently determining whether a claim is true or false. This design separates classification, feature attribution, and natural-language explanation.

Research Gap

The reviewed literature shows three broadly connected research directions. Traditional ML approaches provide useful computational baselines [8], while BERT-based approaches improve contextual text representation [1], [2]. Explainable-AI research addresses the interpretability problem by identifying influential features [3], and recent LLM research provides opportunities for generating natural-language explanations but also highlights concerns regarding hallucination and factuality [7].

Table 1. Research Gaps

Study	Dataset	Model	XAI	LLM	Main Focus
Study 1 [1]	Fake-news dataset	BERT + LightGBM	No	No	Hybrid classification
Study 2 [2]	Fake-news datasets	BERT	No	No	BERT fine-tuning/generalization
Study 3 [3]	Multiple datasets	ML/DL/Transformer	Yes	No	Explainable fake-news detection
Study 4 [4]	Fake-news datasets	Reasoning-based model	Yes	No	Explainable reasoning
Study 5 [7]	⁷ misinformation datasets	Multiple LLMs	Explanation generation	Yes	LLM-based misinformation detection
This Study	LIAR-PLUS	TF-IDF + LR/SVM, DistilBERT, BERT	SHAP + LIME	Controlled local LLM	Classification + interpretation + robustness

The positioning indicates that the present implementation combines comparative classification, local feature attribution, feature-deletion analysis, prediction stability, and controlled LLM-based interpretation within one experimental workflow. The purpose is not to claim that individual components have not previously been studied, but to evaluate their behaviour together in the implemented framework.

Proposed Hybrid Framework

The proposed framework integrates misinformation classification, post-hoc explainability, robustness analysis, and controlled LLM-based interpretation into a single experimental pipeline. The framework uses LIAR-PLUS as the primary dataset and evaluates traditional machine-learning and Transformer-based models before applying SHAP/LIME explanations to the BERT classifier. The resulting model-derived evidence is then provided to a controlled local LLM for generating a concise human-readable explanation.

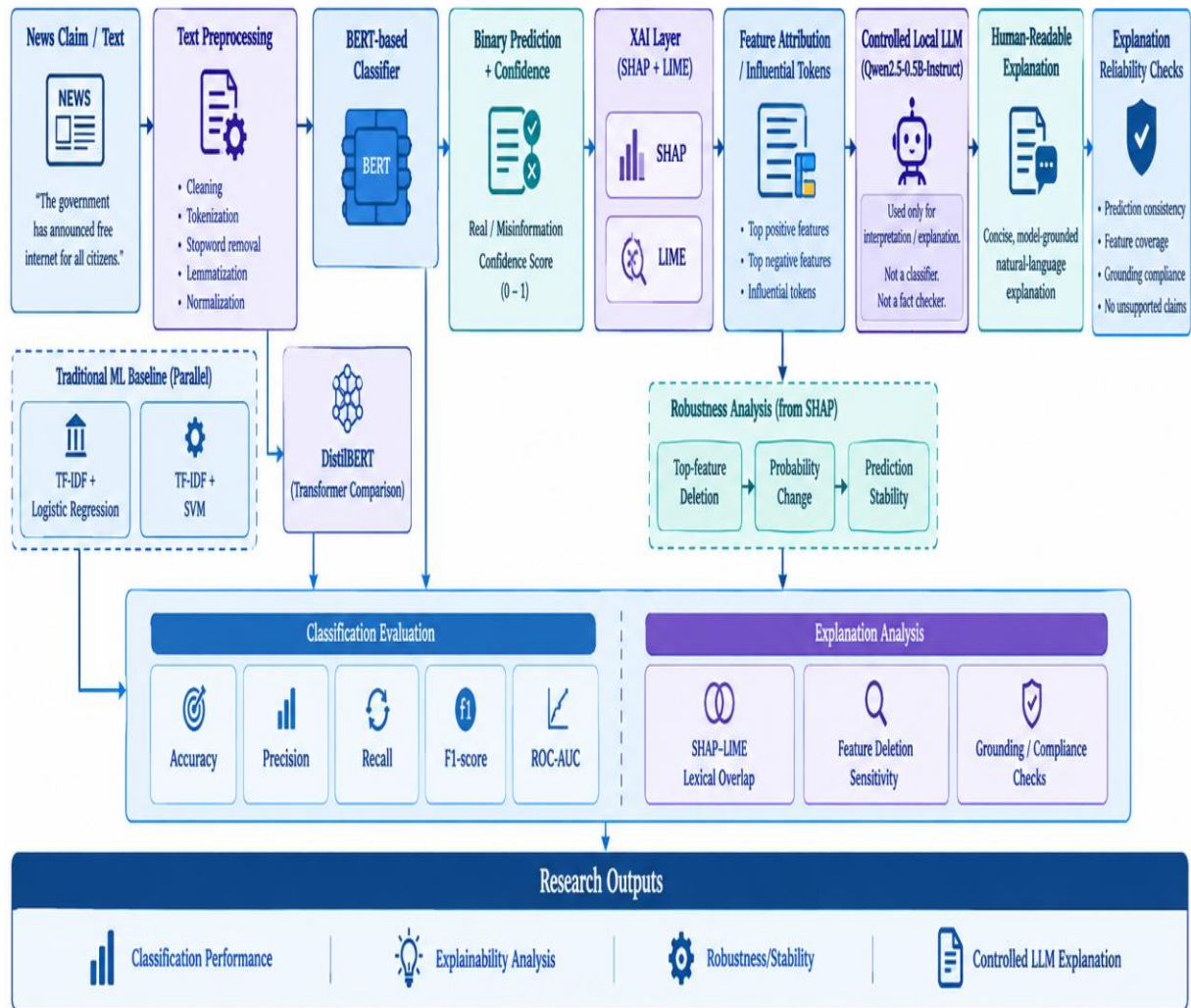


Figure 1. Proposed BERT–XAI–LLM Research Framework

Dataset Input

The study uses the **LIAR-PLUS dataset**, containing short political statements together with associated contextual and justification information. The statement and available textual information are prepared as the input for misinformation classification.

Text Pre-processing

The textual data are processed before model training. The pre-processing stage includes cleaning, tokenization, removal of unnecessary textual noise, and normalization. The resulting text is used for both traditional machine-learning and Transformer-based experiments.

Table 2. Dataset Statistics

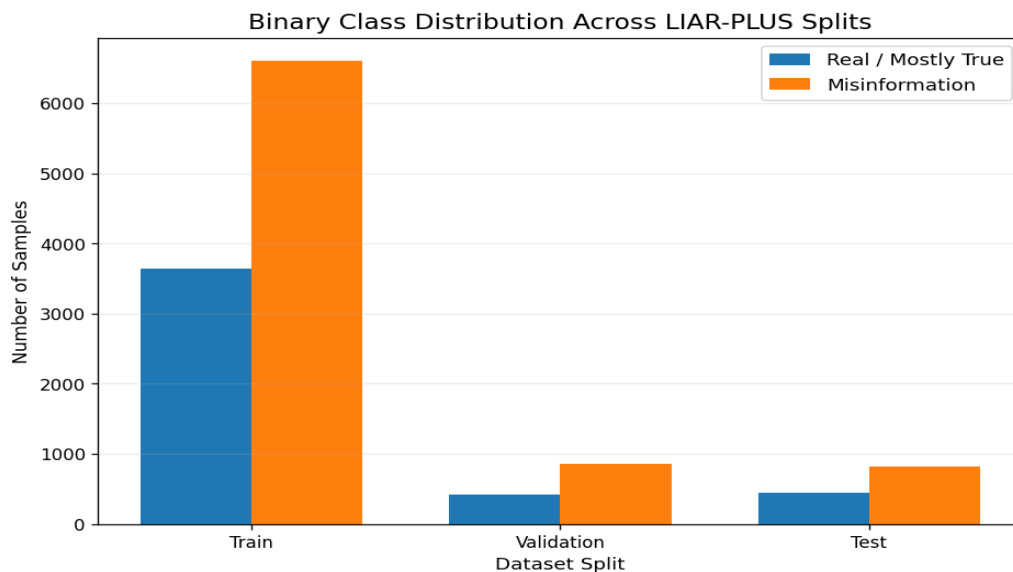
Split	Real	Misinformation	Total
Train	3,638	6,602	10,240
Validation	420	864	1,284
Test	449	818	1,267

Binary Label Transformation

The original LIAR-PLUS labels are converted into two classes for the experimental task:

Table 3. Binary Classification

Original Label	Binary Class
true	Real (0)
mostly-true	Real (0)
half-true	Misinformation (1)
barely-true	Misinformation (1)
mostly-false	Misinformation (1)
false	Misinformation (1)
pants-fire	Misinformation (1)

**Figure 2.** Binary Class Distribution

Classification Models

To evaluate misinformation classification from both conventional and contextual perspectives, four experimental models are implemented in two groups: traditional machine-learning models and Transformer-based models. The same binary classification setting is used for all models, with Real = 0 and Misinformation = 1.

Traditional Machine-Learning Models

TF-IDF + Support Vector Machine (SVM):

The TF-IDF representation converts the textual input into numerical feature vectors based on term importance. SVM is then used to identify a decision boundary between the Real and Misinformation classes. This model serves as a conventional text-classification baseline.

TF-IDF + Logistic Regression (LR):

Logistic Regression is applied to the same TF-IDF feature representation. It estimates the probability of the two classes and provides a computationally efficient baseline for comparison with Transformer-based models.

Transformer-Based Models**DistilBERT**

DistilBERT is used as a lightweight Transformer-based model. It provides contextual text representations while requiring fewer computational resources than the full BERT architecture. It is included to examine whether a smaller Transformer can provide competitive misinformation-classification performance.

BERT

BERT is used as the primary Transformer classifier in the proposed framework. Its bidirectional contextual representation enables the model to consider the surrounding textual context when making a classification decision. The resulting BERT predictions, probabilities, and confidence scores are subsequently passed to the SHAP/LIME explainability layer.

Table 4. Experimental Configuration

Component	Configuration
TF-IDF	10,000 max features
n-gram	(1,2)
SVM	TF-IDF features
Logistic Regression	TF-IDF features
DistilBERT	distilbert-base-uncased
BERT	bert-base-uncased
Max sequence length	128
BERT training	1 epoch
BERT training subset	2,000 samples
Batch size	8
Learning rate	2×10^{-5}

SHAP-Based Explainability

SHAP (SHapley Additive exPlanations) was used to analyse the contribution of individual textual features to the predictions generated by the BERT classifier. For each selected input, SHAP assigns an attribution value to the identified features, indicating their contribution toward or away from the Misinformation class.

In the implemented analysis, a 20-sample subset was examined to obtain local feature-level explanations. A positive SHAP value indicates that the corresponding feature contributed toward the model's Misinformation prediction, whereas a negative SHAP value indicates a contribution away from the Misinformation prediction.

For example, in one analysed instance, features such as “Building,” “wall,” and “Mexico” contributed positively toward the Misinformation score, while features such as “years” contributed negatively. These attributions describe the behaviour of the trained classifier and should not be interpreted as evidence that the underlying claim is factually true or false.

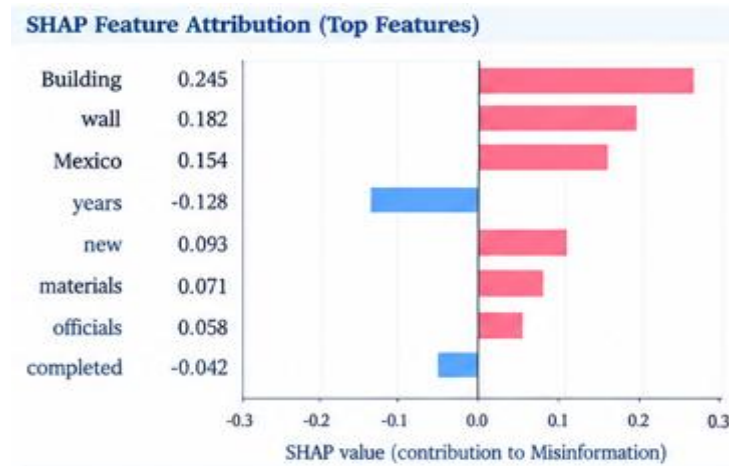


Figure 3. Example of local SHAP feature attribution for a BERT misinformation prediction

LIME-Based Explainability

LIME was used as a complementary local explanation method for examining the behaviour of the BERT classifier. LIME constructs a local surrogate representation around an individual input and identifies textual features associated with the model's prediction.

The same 20-sample subset used for SHAP analysis was also examined using LIME. For each sample, the top contributing textual features were extracted and compared with the corresponding SHAP features.

For example, for the selected sample, LIME identified influential terms including “years” and “Building”, providing a local feature-level interpretation of the model prediction. These explanations were subsequently compared with SHAP using lexical feature overlap.

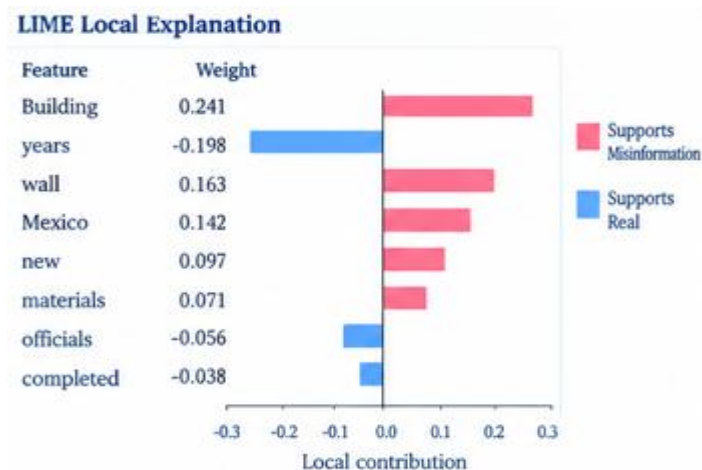


Figure 4. Example of local LIME explanation for the BERT misinformation classifier

SHAP–LIME Comparison

To examine the consistency between the two local explanation methods, the top features identified by SHAP and LIME were compared across the same 20-sample subset. The comparison was performed using Jaccard similarity and the number of common features.

The analysis produced a mean Jaccard similarity of 0.2811, with an average of 1.95 common features per sample. This indicates that SHAP and LIME identified some overlapping influential features, but their top-feature selections were not highly identical across all samples.

- Mean Jaccard similarity = 0.2811
- Mean common features = 1.95

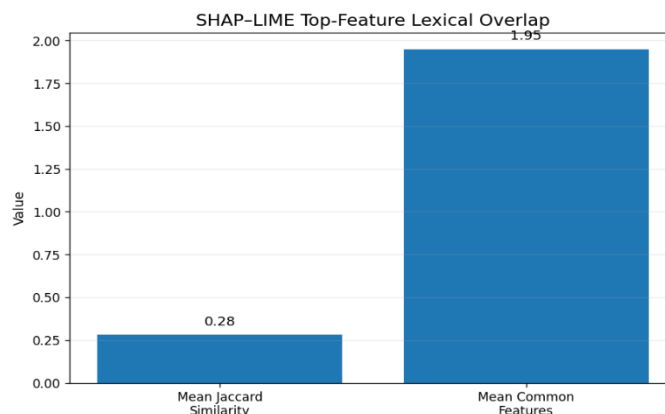


Figure 5. SHAP-LIME Top-Feature Lexical Overlap

SHAP Feature Deletion Analysis

To examine the sensitivity of the BERT predictions to the features identified by SHAP, the top SHAP features were progressively removed from the input text. The model probability was then recalculated and the resulting perturbation was measured using the absolute change in prediction probability.

The analysis was performed by removing the **top-1, top-3, and top-5 SHAP features** from the selected samples. The mean absolute probability change increased from **0.0225** for Top-1 deletion to **0.0471** for Top-3 and **0.0566** for Top-5 deletion.

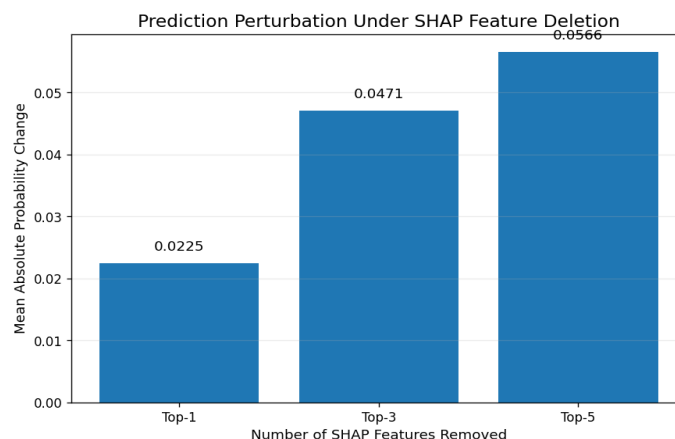


Figure 6. Prediction Perturbation under SHAP Feature Deletion

The increasing probability change with a larger number of removed features indicates that the model's output was progressively perturbed when more SHAP-identified features were removed. This provides an empirical **sensitivity analysis** of the identified features, rather than treating SHAP attribution alone as proof of explanation correctness.

Prediction Stability

Prediction stability was evaluated to determine whether the BERT classification remained unchanged after removing the top SHAP-identified features. Among the 20 analysed samples, 17 predictions remained unchanged, while 3 predictions changed after feature deletion.

This corresponds to an overall prediction stability of 85%. The result indicates that most predictions remained stable under the applied perturbation, while a smaller proportion showed sensitivity to the removal of influential features.

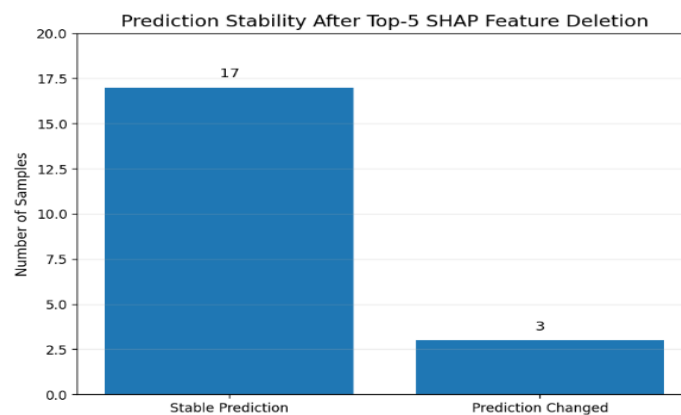


Figure 7. Prediction Stability After Top-5 SHAP Feature Deletion

Controlled LLM Explanation

The BERT prediction, confidence score, and SHAP-derived positive and negative features were provided to Qwen2.5-0.5B-Instruct through a constrained prompt. The LLM was instructed to describe only the supplied model evidence and not to independently fact-check the claim, introduce external information, or state that the prediction was correct or incorrect.

A subset of 20 samples was analysed to examine the behaviour of the controlled LLM. The results indicate that the constrained prompting approach can convert numerical SHAP evidence into a more readable explanation, but the small local LLM still requires careful output validation. In particular, the analysis highlights the importance of checking feature preservation, grounding, and unsupported interpretations before using generated explanations.

Table 5. Controlled LLM Explanation

Sample	Prediction	SHAP Evidence	LLM Explanation	Observation
0	Misinformation	Building, wall, Mexico	Controlled explanation based on the supplied positive/negative SHAP features	Model evidence preserved
3	Misinformation	Medicare, Advantage, supporting features	Controlled explanation describing the supplied feature contributions	Feature-grounded interpretation
7	Misinformation	Top SHAP features from the sample	Controlled explanation based on the supplied attribution evidence	Evidence preservation examined
18	Real	SHAP features associated with the Real prediction	Controlled explanation based on the supplied prediction and attribution information	Grounding validation required

LLM Safety / Grounding Analysis

The controlled local LLM was evaluated to determine whether its generated explanations remained grounded in the BERT prediction and SHAP-derived evidence. The LLM was provided with the predicted class, confidence score, and positive and negative SHAP features under strict prompt constraints.

The controlled local LLM was able to reproduce the model prediction, confidence, and several SHAP-derived features in natural language. However, the analysis also identified cases in which generated explanations introduced unsupported semantic interpretations or factual statements that were not contained in the supplied SHAP evidence.

These observations were found in several analysed samples, including Samples 1, 3, 4, 8, and 18. Therefore, the LLM output was not treated as automatically reliable merely because the generated explanation was fluent or followed the requested format.

The findings indicate that the LLM should be considered an interpretation component requiring grounding and safety validation, rather than an independent fact-checking system. This also supports the use of strict prompting and post-generation checking when small local LLMs are used for explainable misinformation detection.

Results

Classification Performance

Four classification models were evaluated for binary misinformation classification using Accuracy, Precision, Recall, F1-score, and ROC-AUC. The results are presented in Table 8.

Table 6. Main Classification Results

Model	Accuracy	Precision	Recall	F1	ROC-AUC
TF-IDF + SVM	0.6133	0.6843	0.7445	0.7131	0.6081
TF-IDF + LR	0.6677	0.6853	0.8973	0.7771	0.6493
DistilBERT	0.5517	0.6935	0.5477	0.6120	0.5651
BERT	0.6622	0.6871	0.8753	0.7699	0.6506

The results show that TF-IDF + Logistic Regression achieved an accuracy of 0.6677 and an F1-score of 0.7771, with a recall of 0.8973. The BERT model achieved an accuracy of 0.6622, an F1-score of 0.7699, and a recall of 0.8753. The ROC-AUC values of Logistic Regression and BERT were 0.6493 and 0.6506, respectively.

The TF-IDF + SVM model obtained an accuracy of 0.6133 and an F1-score of 0.7131, while DistilBERT obtained 0.5517 accuracy and 0.6120 F1-score. Overall, the results indicate that the evaluated models produced different precision–recall trade-offs. In particular, the Logistic Regression and BERT models showed relatively high recall for the Misinformation class, while the ROC-AUC values remained around 0.61–0.65 across the evaluated models.

For the subsequent SHAP, LIME, feature-deletion, prediction-stability, and controlled LLM analysis, BERT was used as the primary Transformer classifier. This provides a consistent prediction source for the explainability and downstream interpretation stages of the proposed framework.



Figure 8. Overall Model Performance

Explainability Results

The explainability analysis was conducted using SHAP and LIME on the same 20-sample subset of BERT predictions. In addition, SHAP-based feature deletion was performed to examine prediction

sensitivity, while prediction stability was evaluated by comparing the original and perturbed predictions. The results are summarized in Table 9.

Table . Explainability and Prediction Stability Results

Analysis	Result
SHAP–LIME mean Jaccard similarity	0.2811
SHAP–LIME mean common features	1.95
Top-1 deletion mean absolute probability change	0.0225
Top-3 deletion mean absolute probability change	0.0471
Top-5 deletion mean absolute probability change	0.0566
Prediction stability	85%

The mean Jaccard similarity between the top SHAP and LIME features was 0.2811, with an average of 1.95 common features per sample. This indicates that the two explanation methods identified some overlapping influential terms, although their top-feature selections were not highly identical across the analysed samples.

The SHAP-based deletion analysis showed an increase in the mean absolute probability change as more top-ranked features were removed. The mean absolute change increased from 0.0225 for Top-1 deletion to 0.0471 for Top-3 deletion and 0.0566 for Top-5 deletion. This indicates increasing prediction perturbation with larger feature deletions. However, these results represent prediction sensitivity to feature removal and should not be interpreted as direct proof of explanation faithfulness.

The prediction-stability analysis showed that 17 out of 20 predictions remained unchanged, corresponding to 85% stability, while 3 predictions changed after Top-5 SHAP feature deletion. This provides an empirical indication of the extent to which the classifier's predicted class remained stable under the applied feature perturbation.

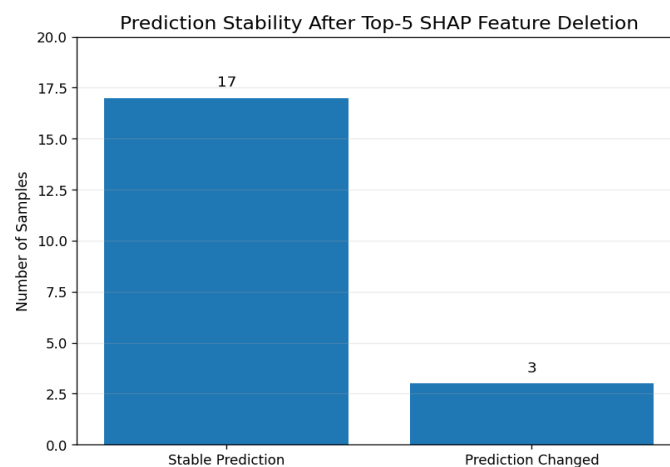


Figure . Prediction stability chart

Discussion

Traditional ML vs. Transformer

The classification results indicate that the traditional TF-IDF-based Logistic Regression model and the BERT model produced relatively close overall performance in the evaluated experimental setting. This suggests that contextual Transformer representations did not provide a large performance difference over the TF-IDF-based baseline under the current training configuration. The relatively close F1-scores also indicate that both approaches were able to capture useful textual patterns for binary misinformation classification.

However, the BERT experiment was conducted using a stratified subset of 2,000 samples and one training epoch on a CPU-based environment. Therefore, the observed results should not be interpreted

as evidence that traditional machine learning is generally superior to BERT. A direct comparison would require comparable training conditions, particularly with respect to dataset size, training duration, computational resources, and hyperparameter optimization. In the present study, BERT was primarily selected as the classifier for the subsequent explainability and LLM-based interpretation pipeline.

Explainability Findings

The explainability analysis demonstrates the complementary nature of SHAP and LIME for interpreting individual BERT predictions. SHAP provides feature-level attribution by estimating the contribution of individual textual features to the model output, whereas LIME constructs a local surrogate representation to approximate the classifier's behaviour around a particular instance.

The observed overlap between the features identified by the two methods was limited, indicating that SHAP and LIME should not be expected to produce identical explanations. This difference can arise from the distinct mechanisms used by the two approaches. Therefore, agreement between SHAP and LIME should be interpreted as overlapping evidence at the feature level, rather than as a requirement for explanation validity. Using both methods provides a broader view of the local behaviour of the classifier.

SHAP Deletion Findings

The SHAP feature-deletion analysis showed that removing progressively larger sets of top-ranked features resulted in greater changes in the model's predicted probability. This indicates that the features identified by SHAP can influence the classifier's output and that their removal can perturb the prediction. However, the direction of probability change was not consistent across all instances. In some cases, removing influential features increased the probability, while in others it decreased it. This suggests that the relationship between individual textual features and the final prediction is not uniformly directional.

Consequently, the deletion experiment should be interpreted as a sensitivity and perturbation analysis rather than as definitive evidence of perfect explanation faithfulness. The analysis demonstrates that SHAP-identified features are associated with changes in model behaviour, but it does not establish that every identified feature is independently necessary or that the explanation completely represents the classifier's decision process.

LLM-Based Explanation Findings

The controlled Qwen2.5-0.5B-Instruct experiment demonstrated that a small local LLM can be used to transform model-derived information into a structured natural-language explanation. Under the controlled prompt, the model generally followed the requested output structure and was able to preserve the predicted class and confidence information in the analysed outputs.

However, the analysis also revealed limitations in the generated explanations. In some cases, the LLM did not reproduce all SHAP-derived features precisely and introduced semantic interpretations that were not explicitly contained in the supplied evidence. Some outputs also introduced contextual or factual statements that were not provided by the classifier or SHAP analysis.

These observations are important because fluent language generation does not necessarily imply evidence-grounded explanation. The LLM therefore should not be treated as an independent fact-checker or evidence source. Instead, it is more appropriately used as a controlled language-generation layer that converts already available model evidence into human-readable text. Strict prompting, feature preservation, and post-generation grounding checks are consequently important for reducing unsupported interpretations.

Overall, the findings support a separation of responsibilities within the proposed framework: BERT performs classification, SHAP and LIME provide model-oriented explanations, feature deletion evaluates prediction sensitivity, and the LLM converts selected model evidence into natural language subject to grounding and safety validation.

Limitations

Computational Limitation

A major limitation of the study is the computational constraint associated with BERT fine-tuning. The BERT model was trained using a stratified subset of 2,000 samples for one training epoch on a CPU-

based environment. Consequently, the reported Transformer performance reflects the specific experimental configuration rather than the maximum performance that could potentially be obtained through full-dataset training, longer training, and systematic hyperparameter optimization using GPU resources.

Limited Explanation Evaluation

The SHAP, LIME, feature-deletion, prediction-stability, and controlled LLM analyses were conducted on a 20-sample subset. Therefore, these analyses provide local and exploratory evidence about model behaviour and explanation characteristics rather than population-level conclusions. A larger evaluation set would be required to assess explanation consistency, robustness, and LLM grounding more comprehensively.

Binary Label Transformation

The original LIAR-PLUS labels were transformed into two classes, Real and Misinformation, for the classification experiments. Under this transformation, statements labelled true and mostly-true were mapped to Real, while half-true, barely-true, mostly-false, false, and pants-fire were mapped to Misinformation. In particular, assigning the half-true category to Misinformation represents a methodological classification choice that simplifies the original multi-class problem but may combine statements with different degrees of factuality. This transformation should therefore be considered when interpreting the reported classification results.

Small Local LLM

The explanation-generation layer used Qwen2.5-0.5B-Instruct, a lightweight local language model selected for computational accessibility. Its small parameter size makes local execution practical, but it also limits its language-generation and reasoning capabilities compared with larger language models. As observed in the experiments, fluent generation does not guarantee that all generated interpretations remain strictly grounded in the supplied model evidence.

No External Evidence Verification

The current LLM layer does not independently verify the factual correctness of the input claims. Its role is limited to converting classifier predictions and SHAP-derived evidence into natural-language explanations. Consequently, an LLM-generated explanation should not be interpreted as independent fact-checking or external evidence verification. This limitation is particularly relevant when the generated explanation introduces information that is not explicitly contained in the supplied model evidence.

Conclusion

This study presented a hybrid framework for explainable misinformation detection that combines traditional machine-learning models, Transformer-based classification, local explainability methods, feature-based robustness analysis, and controlled LLM-based explanation generation. The experimental framework evaluated TF-IDF-based Logistic Regression and SVM models alongside DistilBERT and BERT for binary misinformation classification.

The study further integrated SHAP and LIME to examine the local behaviour of the BERT classifier. Their feature-level comparison showed that the two methods can identify overlapping influential terms while also producing different local feature selections. SHAP-based feature deletion was additionally used to examine prediction sensitivity, and prediction stability was evaluated following the removal of influential features.

A controlled Qwen2.5-0.5B-Instruct layer was then introduced to convert model-derived evidence into structured natural-language explanations. The analysis showed that the local LLM could preserve important prediction information and generate human-readable explanations, but unsupported semantic or contextual interpretations could also occur. Therefore, the LLM should be considered an interpretation and language-generation component rather than an independent fact-checking mechanism.

Overall, the findings demonstrate a practical pipeline in which classification, explainability, robustness analysis, and natural-language generation are treated as separate but connected components. The

results also highlight the importance of grounding and safety validation when LLMs are used to communicate machine-learning model outputs.

Future Work

Several directions can extend the present framework:

- Full-data BERT training: Train and optimize BERT using the complete dataset with GPU resources and more extensive hyperparameter tuning.
- External validation: Evaluate the framework on additional misinformation datasets, such as FakeNewsNet, to examine its generalizability across datasets and domains.
- Larger language models: Investigate larger and more capable open-source LLMs for generating model-grounded explanations.
- Retrieval-Augmented Generation: Integrate Retrieval-Augmented Generation (RAG) to provide access to external evidence while maintaining a clear separation between classifier predictions and retrieved information.
- Automated grounding evaluation: Develop automated measures for feature preservation, unsupported-claim detection, prediction preservation, and evidence alignment.
- Human evaluation: Conduct human-based evaluation of explanation quality, clarity, usefulness, and perceived faithfulness.
- Broader explanation stability testing: Extend SHAP/LIME comparison and feature-perturbation experiments to larger samples and evaluate explanation stability across different instances, models, and perturbation settings.

Declarations

Competing interests. The authors declare that they have no conflict of interest in this research work.

Authors' contributions. All authors participated in the conceptualization, composition, and evaluation of this work, and approved the final version for publication. Name.....andserve as guarantors of this work and thus, assumes responsibility for ensuring the integrity of the offered information.

References

- [1] E. Essa, K. Omar, and A. Alqahtani, "Fake news detection based on a hybrid BERT and LightGBM models," *Complex & Intelligent Systems*, vol. 9, pp. 6581–6592, 2023, doi: 10.1007/s40747-023-01098-0.
- [2] S. Qin and M. Zhang, "Boosting generalization of fine-tuning BERT for fake news detection," *Information Processing & Management*, vol. 61, no. 4, Art. no. 103745, 2024, doi: 10.1016/j.ipm.2024.103745.
- [3] "A systematic survey on explainable AI applied to fake news detection," *Engineering Applications of Artificial Intelligence*, vol. 122, Art. no. 106087, 2023, doi: 10.1016/j.engappai.2023.106087.
- [4] L. Han, X. Zhang, Z. Zhou, and Y. Liu, "A Multifaceted Reasoning Network for Explainable Fake News Detection," *Information Processing & Management*, vol. 61, no. 6, Art. no. 103822, 2024, doi: 10.1016/j.ipm.2024.103822.
- [5] V. Rawte et al., "The Troubling Emergence of Hallucination in Large Language Models—An Extensive Definition, Quantification, and Prescriptive Remediations," *Proceedings of EMNLP*, 2023, pp. 2541–2573.
- [6] Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. N. Georgiev, R. J. Das, and P. Nakov, "Factuality of Large Language Models: A Survey," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.
- [7] H. Wan, S. Feng, Z. Tan, H. Wang, Y. Tsvetkov, and M. Luo, "DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection," *Findings of the Association for*

Computational Linguistics: ACL 2024, pp. 2637–2667, 2024, doi: 10.18653/v1/2024.findings-acl.155.

- [8] H. Alamleh, “Social Media Fake News Detection Using Machine Learning Models and Feature Extraction Techniques,” 2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 134–139, 2024, doi: 10.1109/UEMCON62879.2024.10754749.