

Detecting Misinformation on Social Media Using Transformer-Based Language Models and Explainable Artificial Intelligence

Ayman Aboalndr Mohamed Aboalndr¹ and Waleed M. Ead²✉

¹ Department of Management Information Systems, Higher Institute of Management and Information Technology,
Kafir El- Sheikh, Egypt

² Faculty of Computing and Information, Al-Baha University, Saudi Arabia

dr.aboalndr@himat-kfs.edu.eg; waleed.m@bu.edu.sa

Corresponding Author: waleed.m@bu.edu.sa

Abstract

The rapid dissemination of short claims on social media makes it challenging to detect misinformation, since it involves a classification problem where the model is required to distinguish various levels of truthfulness rather than just true and false. This paper presents an explainable transformer-based model for classifying social-media-style claims that decouples model explanation from factual verification. The public LIAR benchmark that this study revolves around contains a total of 12,836 manually fact-checked and labelled short statements from the source PolitiFact. It uses six different labels for truthfulness. A transformer encoder transforms any claim into a contextual representation, a multiclass classification head estimates the label probabilities, with a token-level explainability layer identifies words with the strongest influence on the prediction. Faithfulness is the reduction in predicted-class score after removing tokens that are highly attributed. The authors also offer a reproducible experimental protocol for BERT and RoBERTa while anchoring the task's anticipated difficulty to released LIAR results, instead of inventing unexecuted measures. The literature demonstrates that six-class LIAR classification is considerably harder than binary fake-news detection, and that quality of the explanation cannot be inferred from accuracy alone. As a result, the proposed framework views the output of the transformer as a triage signal: it can indicate patterns associated with previously fact-checked claims and reveal the text that the classifier uses, but it does not by itself prove that a new claim is true or false. The effective use of automated misinformation detection on social networks and fact-checking workflows hinges on this distinction.

Article History

Manuscript Received
July 07, 2025

First Revision
November 09, 2025

Final Acceptance
November 28, 2025

Keywords: Misinformation Detection; Social Media; Transformer; BERT; RoBERTa; Explainable Artificial Intelligence; LIAR Dataset; Fake News Detection

1 Introduction

On social media, very little time may separate an event, a claim about it, and its exposure to a large audience. The same infrastructure that makes information easy to share can also carry inaccurate or deliberately misleading claims faster than professional verification can respond. Fake-news research therefore sits at the intersection of natural-language processing, social computing, journalism, and platform governance. Shu et al. described this setting as a data-mining problem in which textual content and social context can both contribute useful signals [1].

Automated detection has moved from hand-crafted lexical features toward neural language models. BERT introduced deep bidirectional contextual representations that can be fine-tuned for classification with relatively small task-specific heads [2]. In misinformation research, FakeBERT demonstrated the value of combining BERT representations with deep classification layers for social-media fake-news detection [3]. More recent evaluations have continued to find that pretrained BERT-family representations can be effective for misinformation and rumor classification, although performance varies sharply with dataset construction and task definition [4].

LIAR benchmark renders a particularly tough challenge. Wang developed LIAR using 12,836 short statements that were manually verified for truthfulness. Wang opted to keep the six PolitiFact truthfulness categories rather than collapsing them into a binary label [5]. The dataset comes with an official split of 10,269 training, 1,284 validation, and 1,283 test instances. Descriptions of legal claims only give a part of the legal picture. Often they are also short, requiring more knowledge than what is found in the wording alone. So the dataset is a useful reminder that linguistic classification does not mean verification of the facts. A model can learn patterns associated with labels without having the external evidence that is necessary to determine whether a new unseen claim is true or false.

Transformers make this distinction more important, not less. Their contextual representations can capture subtle linguistic dependencies, but a confident class probability does not reveal which parts of a claim drove the decision. Earlier explainable fake-news systems such as DEFEND and GCAN showed that interpretable evidence can be exposed through attention over news sentences, comments, retweeters, and words [6], [7]. A systematic survey of XAI for fake-news detection nevertheless found that explanation methods, audiences, and evaluation criteria remain heterogeneous [8]. Recent user-centered work goes further: Bezzaoui et al. found that adding explanations to a disinformation-detection artifact did not automatically increase trust or usability [9].

This paper explores misinformation detection by looking at two different questions: how a transformer model assigns a fine-grained truthfulness label and how the textual basis of that

assignment can be inspected. The framework proposed within the paper uses a transformer classifier combined with token-level attribution and a faithfulness test that removes the most influential tokens. The explanation provided by the paper is intended to explain the model's behavior rather than to demonstrate the truth or falsity of the claim. The paper is guided by three key questions:

1. How can transformer-based language representations be formulated for the six-class LIAR task?
2. Can token-level XAI identify the relevant lexical evidence for each classification in a faithful manner?
3. What limitations exist in using a text classifier to screen for misinformation without the need to retrieve independent evidence or perform fact-checking?

The primary innovation is an auditable transformer-XAI workflow, a reproducible LIAR protocol, and a careful segregation between predictive explanation and factual verification. The contributions of this research are as follows:

1. A six-class transformer framework that retains the original LIAR truthfulness scale.
2. An explainability layer that reports the token-level evidence provided by the transformer classifier and tests the faithfulness of that explanation through a token deletion test.
3. A dataset-aware protocol that distinguishes the text of the claim from the speaker and the context of the claim so that the effects of each factor on the classification can be examined separately.
4. An evaluation design that examines both the performance of the transformer classifier on a number of tasks and metrics while also examining the faithfulness of the transformer's explanations of its classifications.

2 Background Study

2.1 Transformer Models for Social-Media Misinformation Detection

Computational fake-news research has long used both content and social context. The survey by Shu et al. organized early work around news content, user engagement, and propagation information [1]. This is particularly relevant to social-media claims because the same writing style can occur in both accurate and inaccurate content, making auxiliary context useful when it is available.

Recent systems increasingly combine several sources of evidence. Truică, Apostol, and Karras proposed DANES, which uses separate textual and social branches and reported gains from their combination on BuzzFace, Twitter15, and Twitter16 [10]. Yu and colleagues proposed DEP-FEND, combining historical news-environment signals with external factual evidence; their experiments on two Chinese datasets showed that text-based and fact-checking signals can complement each other [11]. These studies motivate a cautious interpretation of text-only LIAR experiments: linguistic models are useful, but external evidence and social context can add information that is absent from the claim itself.

Recent studies continue to show the value of transformer representations. Reshi and Ali evaluated MPNet, SentenceT5, and GTR embeddings on Fakeddit and reported strong transfer-learning performance for rapid misinformation detection [12]. Anggrainingsih, Hassan, and Datta compared BERT-based language models for misinformation detection and highlighted the effect of pretrained embeddings, data splitting, and classifier design [4].

The value of transformer representations is still evident in recent studies. Reshi and Ali evaluated MPNet, SentenceT5, and GTR embeddings on the Fakeddit dataset and noted the strong performance of these models in detecting misinformation with a transfer learning approach [12]. Anggrainingsih, Hassan, and Datta, explored the potential of BERT-based models for misinformation detection and found that pretrained embeddings, data splitting, and classifier design all played significant roles in the performance of these models [4].

2.2 Explainability and Trust in Misinformation Detection

BERT is particularly well suited to short-claim classification tasks due to its contextual understanding of tokens within sentences [2]. On the LIAR dataset, an evaluation of short text claims revealed the challenge of classifying these text instances using only surface features, with the published hybrid model achieving a score of only 27.4% on the test data [5]. By integrating claim and metadata fields into text in a patterned approach to build a BERT-based mental model for claim classification, researchers like Ding, Hu, and Chang were able to achieve a performance score that was 16.71% better than what the state of the art could achieve at the time of their studies [13].

These results explain why the present paper keeps the original six labels. Converting LIAR to binary classes can yield higher headline accuracy, but it removes the distinction between barely-true, half-true, and mostly-true statements. Fine-grained classification is harder, yet it better reflects the graded judgments in the source dataset [5]. It also makes macro-averaged evaluation important because the pants-fire category is smaller than the other five classes.

Current research is also moving beyond plain text. LekshmiAmmal and Madasamy combined transformer and large-language-model components in an explainable multimodal system for low-resource fake-news detection [14]. Athira A.B., Madhu Kumar, and Chacko used pretrained text and vision transformers with SHAP to analyze modality contributions in multimodal fake-news detection [15]. Bhukya, Shukla, and Singh used RoBERTa and vision-transformer encoders with reliability-aware fusion for social-media fake news [16]. These studies show where the field is heading, but they also make a controlled text-only benchmark valuable for isolating what the language encoder itself learns.

Explainability in misinformation detection serves at least two different purposes. One is diagnostic: researchers want to know whether a model is exploiting sensible evidence or dataset artifacts. The other is communicative: users and fact-checkers need an understandable account of why an item was flagged. dEFEND used sentence-comment co-attention to identify check-worthy sentences and comments [6], while GCAN highlighted suspicious retweeters and words in a social-media setting [7]. Both approaches illustrate that an explanation should be connected to the evidence actually used by the model.

The field has since become more explicit about explanation quality. Athira A.B., Madhu Kumar, and Chacko surveyed XAI approaches for fake-news detection and organized prior work by feature type, explanation type, intended explainee, and evaluation metric [8]. Hashmi and colleagues combined deep fake-news detection with LIME-based explanation, demonstrating the continuing use of local post-hoc explanations in this domain [17]. Bezzaoui et al., however, found that richer explanation features can also introduce uncertainty and do not automatically improve user trust [9].

The proposed framework follows this more restrained view. Token attribution is used to show what influenced a transformer classification, and faithfulness is tested by perturbing the influential tokens. The explanation is not presented as a fact-checking justification. That boundary is important because evidence-aware systems and recent fact-checking research increasingly retrieve external material precisely because textual plausibility is insufficient for establishing truth [11].

3 LIAR Dataset and Attribute Analysis

Each LIAR record contains the claim itself together with contextual information about the speaker and the fact-checking record. Table 1 summarizes the attributes available in the original LIAR release [5].

Table 1. Attributes of the LIAR dataset [5]

Attribute	Type / form	Description	Role in the proposed protocol
ID	Integer / identifier	Unique identifier assigned to each fact-checked statement.	Tracking / split integrity
Label	Categorical, 6 classes	Truthfulness rating: pants-fire, false, barely-true, half-true, mostly-true, or true.	Prediction target
Statement	Text	Short fact-checked claim collected from PolitiFact.	Core transformer input
Subject(s)	Categorical / text	Topic or subject associated with the claim; a statement may have multiple subjects.	Metadata ablation only
Speaker	Categorical / text	Name of the person or source making the statement.	Metadata ablation only

Attribute	Type / form	Description	Role in the proposed protocol
Speaker job title	Categorical / text	Occupation or official role of the speaker when available.	Metadata ablation only
State info	Categorical / text	U.S. state associated with the speaker when available.	Metadata ablation only
Party affiliation	Categorical	Political-party affiliation of the speaker.	Metadata ablation only
Barely-true count	Integer	Number of the speaker's previous statements rated barely-true.	Metadata ablation only
False count	Integer	Number of the speaker's previous statements rated false.	Metadata ablation only
Half-true count	Integer	Number of the speaker's previous statements rated half-true.	Metadata ablation only
Mostly-true count	Integer	Number of the speaker's previous statements rated mostly-true.	Metadata ablation only
Pants-fire count	Integer	Number of the speaker's previous statements rated pants-fire.	Metadata ablation only
Context / venue	Categorical / text	Setting in which the statement was made, such as an interview, speech, debate, or social-media post.	Metadata ablation only

The dataset is distributed with an official split of 10,269 training statements, 1,284 validation statements, and 1,283 test statements. Wang reported an average statement length of 17.9 tokens [5]. In the proposed protocol, the validation split is reserved for model selection and early stopping, and the official test split is used only for final evaluation. The six labels are retained rather than collapsed into a binary task because preserving all six categories keeps the distinction between partially true, mostly true, and clearly false claims.

4 Transformer-XAI Detection Architecture

Figure 1 presents the proposed system architecture. A short claim is tokenized and passed to a pretrained transformer encoder. The contextual representation associated with the classification token is then mapped to six truthfulness probabilities. The XAI layer operates on the same prediction and produces token-level attributions, after which a faithfulness check tests whether removing highly attributed tokens changes the model score.

Transformer-XAI architecture for six-class misinformation detection

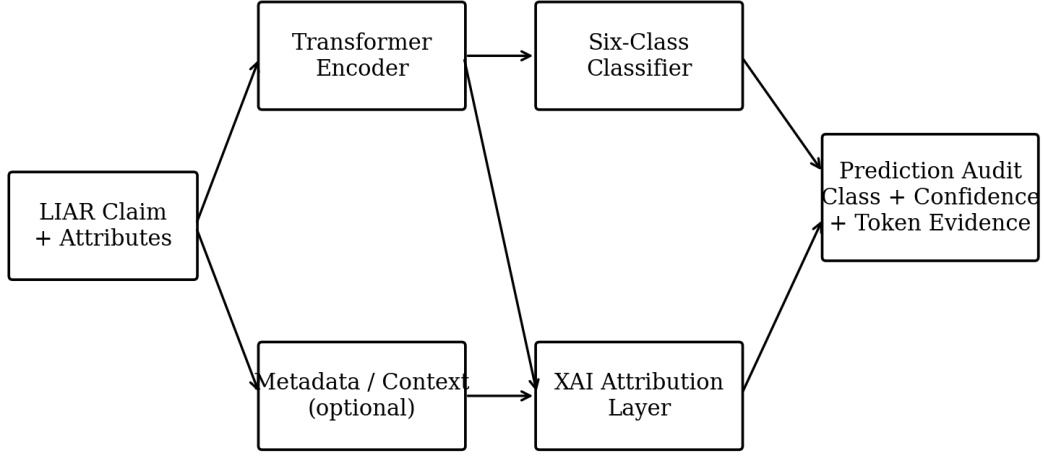


Figure 1. Layered architecture of the proposed Transformer-XAI misinformation detection system.

For an input claim x , the transformer produces a contextual representation:

$$h = \text{Enc}_\theta(x) \quad (1)$$

where Enc_θ denotes the pretrained transformer encoder with parameters θ , x is the tokenized claim, and h is the contextual representation used by the classification head. In BERT-style models, h is normally derived from the special classification token. The encoder parameters may be fine-tuned jointly with the classifier.

The probability assigned to truthfulness class c is obtained with a softmax classification head:

$$p_c = \frac{\exp(\mathbf{w}_c^\top \mathbf{h} + b_c)}{\sum_{j=1}^C \exp(\mathbf{w}_j^\top \mathbf{h} + b_j)} \quad (2)$$

where $C=6$ for LIAR, $\mathbf{w}_c$ and b_c are the class-specific weight vector and bias, and p_c is the predicted probability for class c . The six probabilities sum to one. The predicted label is

$$\hat{y} = \arg \max_{c \in \{1, \dots, C\}} p_c \quad (3)$$

where $\hat{y}$ is the selected truthfulness class. This probability is a model score; it should not be interpreted as a calibrated probability that the real-world claim is factually true.

The model is trained with multiclass cross-entropy:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{ic} \log p_{ic} \quad (4)$$

where N is the number of training instances, y_{ic} is equal to one when training instance i belongs to class c , and zero otherwise, and p_{ic} is the predicted probability that training instance i belongs to class c . This objective function is standard for problems with mutually exclusive class labels, and is appropriate for the six classes in the original LIAR dataset.

To explain the models locally, the framework can utilize SHAP-style token attributions. The models can be written as

$$g(\mathbf{z}') = \phi_0 + \sum_{j=1}^M \phi_j z'_j \quad (5)$$

where $\mathbf{z}'$ is the simplified binary representation indicating whether interpretable token features are present, M is the number of interpretable features, ϕ_0 is the baseline value, and ϕ_j is the attribution assigned to token feature j . Positive and negative attributions indicate how a token moves the explained class score relative to the baseline; they do not establish factual support for the claim.

A local explanation should also be tested rather than accepted because it looks plausible. Let S_k denote the k tokens with the largest absolute attributions. The deletion-based faithfulness score is

$$\Delta_k = f_{\hat{y}}(x) - f_{\hat{y}}(x \setminus S_k) \quad (6)$$

where $f_{\hat{y}}(x)$ is the predicted-class score for the original claim, and $f_{\hat{y}}(x \setminus S_k)$ is the score after removing or masking the top- k attributed tokens from the text. A larger positive value for Δ_k indicates that the highlighted tokens significantly supported the original prediction. It's important to note that this metric evaluates how well an explanation reflects the classifier's decision, not whether the prediction is factually correct.

For the evaluation, the study used the official LIAR train/validation/test partition as described in Section 3 [5]. The core transformer model only used the statement text as input, while speaker information and any contextual metadata were kept separate and reserved for a separate ablation study. Keeping the input features separated in this way makes it easier to determine whether the model's performance change is due to the claim text itself or the additional metadata.

$$\text{MacroF1} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (7)$$

where $F1_c$ is the $F1$ score computed for class c and $C=6$. Accuracy, macro-precision, macro-recall, per-class $F1$, and the confusion matrix should also be reported. Macro averaging prevents the largest classes from dominating the summary score.

The reproducible protocol specifies two transformer baselines: BERT-base-uncased and RoBERTa-base, each fine-tuned on the official training split, selected on validation macro-F1, and evaluated once on the untouched test split. Tokenization uses each model’s native tokenizer. A practical starting configuration is a maximum sequence length of 128, batch size 16, AdamW optimization, learning rate 2×10^{-5} , weight decay 0.01, and up to five epochs with early stopping on validation macro-F1. All random seeds, package versions, model checkpoints, and preprocessing steps should be recorded in the final experimental release. These settings are protocol parameters rather than measured findings and should be tuned only on the validation set.

5 Experimental Evaluation

5.1 Classification Performance and Benchmark Comparison

New transformer results are reported only after the specified checkpoints have actually been fine-tuned and evaluated. In the present manuscript, no unexecuted BERT or RoBERTa result is presented as a new measurement. Table 2 therefore summarizes only numerical results stated in the cited publications, while Table 3 reproduces Wang’s original LIAR benchmark in detail. Wang’s text-and-all-metadata hybrid CNN achieved 27.4% held-out test accuracy [5]. Ding, Hu, and Chang reported a 16.71% accuracy improvement over the state of the art available at the time for their BERT-based mental model [13].

Table 2. Published numerical results on the six-class LIAR benchmark

Published model	Input setting	Published numerical result	Source status
Wang [5], hybrid CNN	Claim + all metadata	27.4% test accuracy	Directly reported
Ding, Hu, and Chang [13], BERT mental model	Claim + metadata	16.71% reported accuracy improvement over prior SOTA	Directly reported

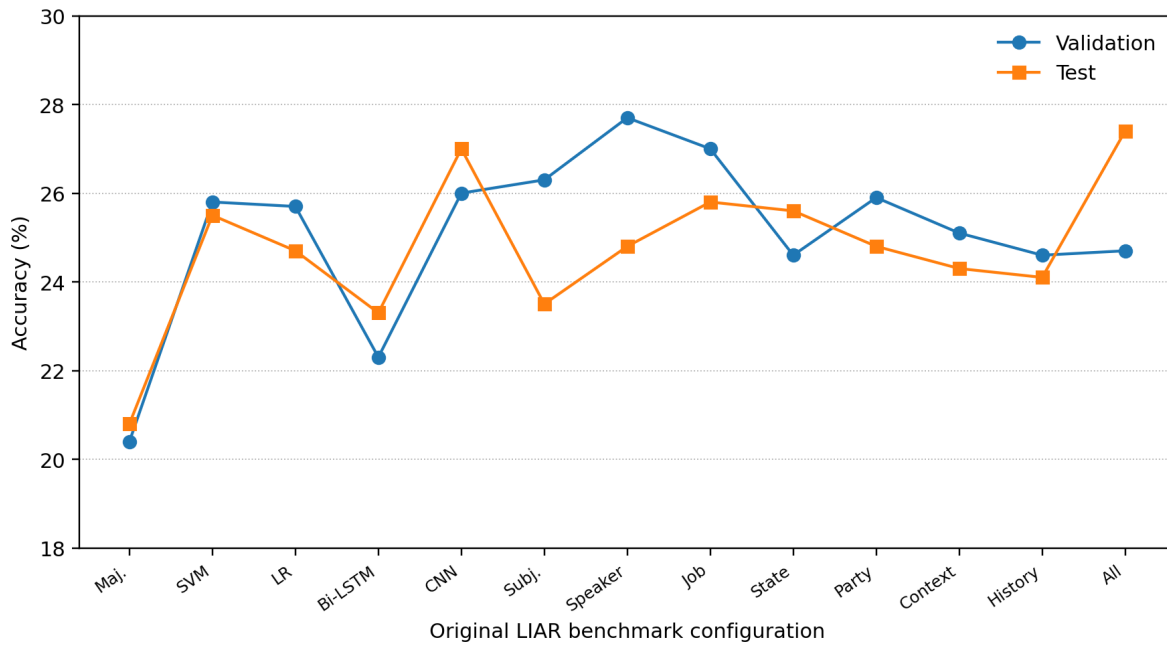


Figure 2. Validation and test accuracy curves for the original LIAR benchmark configurations reported by Wang [5].

To provide a broader numerical reference for the six-class task, Table 3 reproduces the validation and held-out test accuracies reported in the original LIAR benchmark [5]. The final column is calculated from those published test accuracies as the absolute percentage-point gain over the 20.8% majority baseline; it is not a separate metric reported by Wang. The strongest held-out result in the original study was 27.4% for the Text + All hybrid CNN, compared with 27.0% for the text-only CNN and 20.8% for the majority baseline.

Table 3. Original LIAR benchmark accuracies [5] and derived gain over the majority baseline

Model / configuration	Feature setting	Validation accuracy (%)	Test accuracy (%)	Test gain vs. majority (pp)
Majority	Label frequency	20.4	20.8	0.0
SVM	Claim text	25.8	25.5	4.7
Logistic Regression	Claim text	25.7	24.7	3.9
Bi-LSTM	Claim text	22.3	23.3	2.5
CNN	Claim text	26.0	27.0	6.2
Hybrid CNN	Text + Subject	26.3	23.5	2.7
Hybrid CNN	Text + Speaker	27.7	24.8	4.0
Hybrid CNN	Text + Job	27.0	25.8	5.0
Hybrid CNN	Text + State	24.6	25.6	4.8
Hybrid CNN	Text + Party	25.9	24.8	4.0
Hybrid CNN	Text + Context	25.1	24.3	3.5
Hybrid CNN	Text + History	24.6	24.1	3.3
Hybrid CNN	Text + All metadata	24.7	27.4	6.6

The variation across the published results is informative in its own right. LIAR performance is sensitive to whether speaker and contextual metadata are used, how those fields are encoded, and whether the task is kept at six classes or simplified. A fair comparison therefore needs to report the exact input fields and label mapping. Results from binary conversions should not be placed in the same table as the original six-class task without a clear warning.

Table 4. Evaluation metrics used for predictive and explanation assessment

Metric	Scope	Range	Preferred direction
Accuracy	Overall six-class classification	0–1	Higher
Macro-Precision	Equal-weight class precision	0–1	Higher
Macro-Recall	Equal-weight class recall	0–1	Higher
Macro-F1	Equal-weight class F1	0–1	Higher
Deletion faithfulness Δk	Predicted-score drop after masking top-k tokens	Typically –1 to 1	Larger positive drop

5.2 Explanation Faithfulness and Error Analysis

Once the transformer experiments have been completed, the explainability stage can be examined claim by claim. For every test claim, whether classified correctly or not, the framework records the predicted label, probabilities across the six classes, tokens receiving the strongest attribution, and the deletion score Δk . What matters here is whether the highlighted words have a meaningful effect on the model’s decision. An explanation may look convincing visually, but it offers limited insight if removing the most influential tokens produces almost no change in the prediction. For this reason, the analysis considers the distribution of Δk across both correct and incorrect classifications. Particular attention is also given to high-confidence cases where the prediction appears to depend more heavily on names, political terms, or stylistic patterns than on the meaning of the claim itself.

This type of examination is important for reasons already raised in earlier research. Explainable fake-news detection methods can reveal textual or social signals that contribute to a prediction, but providing an explanation does not necessarily make the underlying detector more reliable or more trustworthy to its users [9]. The explanation is therefore better understood as a means of examining what the model has learned, rather than as evidence that its truthfulness judgment must be correct. For example, if predictions repeatedly depend on speaker identities, partisan language, or common lexical shortcuts, the explanation exposes a weakness in the classifier. In such cases, it should not be interpreted as support for deploying the model.

The system is consequently intended mainly as a triage tool in practical settings. The transformer can prioritize or flag claims whose language resembles patterns learned from previously fact-checked misinformation, while the accompanying explanation gives a reviewer some indication of what drove the prediction. Determining whether a claim is actually true or false remains a different

task and requires supporting evidence, evaluation of the source, or professional fact-checking. Evidence-aware approaches such as DEP-FEND reinforce this separation by demonstrating how external factual evidence can supplement information obtained from the claim text itself [11].

6 Conclusion

This paper develops an explainable transformer framework for fine-grained misinformation detection on social-media-style claims. The design retains the original six LIAR truthfulness categories and combines a contextual language encoder with a multiclass classifier, token-level attribution, and a deletion-based faithfulness test. The resulting explanation is intended to show which words affected the classifier rather than to manufacture a factual justification after the prediction has been made.

Published benchmark results confirm that LIAR remains a difficult task and that reported performance changes substantially with the input representation and use of metadata. Wang's original benchmark and the later BERT-based mental model report markedly different six-class accuracies. This variation makes it important to disclose the exact label mapping, input fields, and evaluation split. It also supports the decision to use macro-F1 alongside accuracy when the planned BERT and RoBERTa experiments are executed.

Several limitations remain. LIAR is centered on short political claims and PolitiFact judgments, so its distribution does not represent all forms of social-media misinformation. Text-only classification cannot independently verify a claim against current events or external evidence, and token attribution explains model sensitivity rather than real-world truth. The present document also does not report invented transformer measurements: the proposed BERT/RoBERTa runs require the actual pretrained checkpoints and execution environment before their values can be entered. This limitation is preferable to presenting plausible but unverified numbers as experimental evidence.

A natural next step is to execute the fixed transformer protocol, compare BERT and RoBERTa under identical splits, and evaluate explanation faithfulness quantitatively and with human fact-checkers. A second stage should add evidence retrieval so that linguistic classification can be separated from factual support. Multimodal content, multilingual claims, temporal drift, adversarial robustness, calibration, and user-centered explanation design are also important directions. Recent multimodal and explainable transformer studies indicate that these extensions are technically feasible, but the central requirement should remain the same: the system must make clear whether it is explaining a classifier or verifying a fact.

7 References

- [1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017. doi: 10.1145/3137597.3137600.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.
- [3] R. K. Kaliyar, A. Goswami, and P. Narang, “FakeBERT: Fake news detection in social media with a BERT-based deep learning approach,” *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11765–11788, 2021. doi: 10.1007/s11042-020-10183-2.
- [4] R. Anggrainingsih, G. M. Hassan, and A. Datta, “Evaluating BERT-based language models for detecting misinformation,” *Neural Computing and Applications*, vol. 37, pp. 9937–9968, 2025. doi: 10.1007/s00521-025-11101-z.
- [5] W. Y. Wang, “Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 422–426, 2017. doi: 10.18653/v1/P17-2067.
- [6] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, “dEFEND: Explainable Fake News Detection,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 395–405, 2019. doi: 10.1145/3292500.3330935.
- [7] Y.-J. Lu and C.-T. Li, “GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 505–514, 2020. doi: 10.18653/v1/2020.acl-main.48.
- [8] Athira A. B., S. D. Madhu Kumar, and Anu Mary Chacko, “A systematic survey on explainable AI applied to fake news detection,” *Engineering Applications of Artificial Intelligence*, vol. 122, 106087, 2023. doi: 10.1016/j.engappai.2023.106087.
- [9] I. Bezzaoui, C. Stein, C. Weinhardt, and J. Fegert, “Explainable AI for online disinformation detection: Insights from a design science research project,” *Electronic Markets*, vol. 35, article 66, 2025. doi: 10.1007/s12525-025-00799-3.
- [10] C.-O. Truică, E.-S. Apostol, and P. Karras, “DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection,” *Knowledge-Based Systems*, vol. 294, 111715, 2024. doi: 10.1016/j.knosys.2024.111715.
- [11] W. Yu, J. Ge, Z. Chen, H. Liu, M. Ouyang, Y. Zheng, and W. Kong, “Research on Fake News Detection Based on Dual Evidence Perception,” *Engineering Applications of Artificial Intelligence*, vol. 133, 108271, 2024. doi: 10.1016/j.engappai.2024.108271.
- [12] J. A. Reshi and R. Ali, “Defending against Misinformation: Evaluating Transformer Architectures for Quick Misinformation Detection on Social Media,” *Procedia Computer Science*, vol. 235, pp. 2909–2919, 2024. doi: 10.1016/j.procs.2024.04.275.
- [13] J. Ding, Y. Hu, and H. Chang, “BERT-Based Mental Model, a Better Fake News Detector,” in *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence (ICCAI '20)*, pp. 396–400, 2020. doi: 10.1145/3404555.3404607.
- [14] H. R. LekshmiAmmal and A. K. Madasamy, “A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers,” *Journal of Big Data*, vol. 12, article 46, 2025. doi: 10.1186/s40537-025-01093-x.
- [15] Athira A. B., S. D. Madhu Kumar, and Anu Mary Chacko, “Pretrained transformers for multimodal fake news detection: Explainability using SHapley Additive exPlanations for contributions from text, image, and image captions,” *Engineering Applications of Artificial Intelligence*, vol. 162, 112590, 2025. doi: 10.1016/j.engappai.2025.112590.

- [16] R. K. Bhukya, A. Shukla, and N. Singh, "Multimodal social media fake news detection using RoBERTa and vision transformer encoders with reliability aware adaptive fusion," *Discover Computing*, vol. 29, article 405, 2026. doi: 10.1007/s10791-026-10281-z.
- [17] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali, and M. Abomhara, "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI," *IEEE Access*, vol. 12, pp. 44462–44480, 2024. doi: 10.1109/ACCESS.2024.3381038.