Jurnal The Messenger, Vol. 18, No. 1 (2026), pp. 120-136
P-ISSN: 2086-1559, E-ISSN: 2527-2810
DOI: 10.26623/themessenger.v18i1.16007

Explainable Stacked Ensemble Learning for Fake News Detection using Text and Metadata

K. M. Tousif Bin Parves[1], Prashanta Kumar Shill[2]
[1]ORCID iD: 0009-0006-9309-879X, Port City International University, South Khulshi, Chattogram 4225, Bangladesh
[2]ORCID iD: 0009-0002-8255-6995, Port City International University, South Khulshi, Chattogram 4225, Bangladesh
*Corresponding author, e-mail: prashanta@portcity.edu.bd

Abstract
Purpose: Automated fake news detection can support verification, but predictions are less useful when textual and contextual evidence cannot be inspected. The study examined whether combining claim text with speaker metadata could improve veracity classification while retaining explanations for journalistic review.
Methods: The LIAR train, validation, and test partitions were used, with six veracity labels remapped into Fake and Real classes. A BERT-base text classifier and Extreme Gradient Boosting metadata classifier produced class probabilities combined by an L1-regularized logistic regression meta-learner with disagreement and interaction features. SHapley Additive exPlanations, Local Interpretable Model-agnostic Explanations, and Integrated Gradients inspected metadata, ensemble, and token-level behavior.
Findings: The stacking model reached 0.7514 accuracy, 0.7483 macro-F1, 0.8288 area under the receiver operating characteristic curve, and 0.4970 Matthews correlation coefficient. It exceeded the BERT text branch (macro-F1=0.6203) and metadata branch (macro-F1=0.7066). Metadata provided the stronger signal, with credibility score and speaker-history variables receiving the largest SHAP importance. High text-metadata disagreement occurred in 30.4% of test cases, and the ensemble reached 0.8234 accuracy within this subset.
Originality: The study extends explainable fake-news detection by conceptualizing disagreement between evidence sources as a dimension of interpretability, separating semantic claim evidence, contextual speaker-history evidence, and their interaction at the fusion stage.
Keywords: Fake News Detection, Explainable Artificial Intelligence, Stacked Ensemble Learning, Text-Metadata Disagreement, Speaker Credibility.

Introduction

The rapid dissemination of false and misleading information has become a major problem for journalism, public communication, and democratic discourse. Disinformation does not only create isolated factual errors; repeated exposure can alter trust among citizens, media organizations, and political institutions. Communication studies have connected the present information disorder with declining trust in news, polarization, and the difficulty of maintaining evidence-based public discussion. The problem becomes more visible on online platforms, where distribution depends on algorithms, interaction patterns, and emotional response rather than editorial review alone. Under these conditions, manual verification becomes difficult to scale, particularly when questionable claims circulate faster than corrections (Bennett & Livingston, 2018; Lazer et al., 2018; Ognyanova et al., 2020; Pennycook & Rand, 2021; Tandoc Jr. et al., 2018).

For journalism, the difficulty is technical, professional, and ethical at the same time. Studies of artificial intelligence in journalism show that automated systems are

Article History: *Received April 02, 2026; Revised August 26, 2026; Accepted September 04, 2026; Published September 09, 2026.*

increasingly considered for monitoring, search, filtering, attribution, and verification, while journalists remain concerned about tools that provide decisions without usable reasons. Acceptance is therefore connected not only with predictive performance but also with transparency, accountability, and the possibility of human inspection. A classification result that cannot be questioned is of limited value in a newsroom, where editorial responsibility remains with the journalist rather than with the model (Broussard et al., 2019; Calvo-Rubio & Ufarte-Ruiz, 2021; Gutiérrez-Caneda et al., 2024; Luttrell et al., 2025; Palla & Kostarella, 2025).

Automated fake news detection has consequently moved from manually designed lexical and stylistic features toward neural representation learning, transformer models, and hybrid architectures. Early systems showed that linguistic cues could support classification, but their behavior was sensitive to the dataset, topic, and type of deception being studied. Deep learning later improved contextual representation, while transformer-based models made it possible to model richer semantic relations in short claims and longer news content. Even so, survey work continues to show that performance depends strongly on benchmark construction, annotation practice, domain, and available context, so aggregate scores should be interpreted together with the characteristics of the data being used (Collins et al., 2021; Hu et al., 2025; Sharma & Singh, 2024; Zhou & Zafarani, 2020).

Text-based deep learning has become a common direction because misinformation can be reflected in framing, wording, stance, semantic inconsistency, and rhetorical choice. BERT-based and hybrid architectures have produced useful results across misinformation benchmarks, yet textual content alone does not always contain enough evidence for a reliable veracity decision. This is particularly relevant to political claims, where the identity and previous record of a speaker can supply contextual information that is absent from a short sentence. For this reason, hybrid systems that combine content with structured or relational information have received increasing attention (Alghamdi et al., 2024; Aslam et al., 2025; Kaliyar et al., 2021; Koloski et al., 2022; Kula et al., 2022; Nasir et al., 2021; Rai et al., 2022; Shu et al., 2020).

The LIAR benchmark offers a controlled setting for examining this combination. It contains short fact-checked political statements collected from PolitiFact together with speaker and contextual metadata, and the original task uses six truthfulness labels: pants-fire, false, barely-true, half-true, mostly-true, and true (W. Y. Wang, 2017). These labels represent degrees of factual accuracy rather than six unrelated forms of news. In a binary setting, the first three can be treated as Fake and the remaining three as Real, making it possible to compare semantic claim evidence with structured speaker-history evidence while still retaining the original labels for later error analysis.

The use of metadata changes the interpretation problem. Text modeling reflects the wording and meaning of the claim, whereas metadata can encode party, state, occupation, previous fact-check history, and derived indicators of past credibility. When these sources point in the same direction, a final decision is comparatively easy to understand. When they point in different directions, however, the analyst needs to know whether the final prediction is mainly supported by the text, by the contextual record, or by an interaction between the two. Hybrid performance therefore introduces a second interpretability problem at the fusion stage (Chalehchaleh et al., 2024; De Magistris et al., 2022; Hashmi et al., 2024; Nwaiwu et al., 2025).

Explainable artificial intelligence provides several tools for examining this problem, but explanation quality depends on the model and the type of evidence. Global

feature attribution can be informative for structured variables, while local surrogate methods and token attribution are more suited to individual textual predictions. General explainability studies distinguish between explanations that are visually plausible and explanations that remain faithful to the learned decision function. Reviews of fake news and automated fact verification similarly argue that explanations should be assessed for fidelity, stability, and usefulness rather than being presented only as decorative salience maps (Adadi & Berrada, 2018; Athira et al., 2023; Guidotti et al., 2018; Mishima & Yamana, 2022; Vallayil et al., 2023).

Recent empirical work has started to combine misinformation detection with SHAP-based attribution, feature-driven explanation, stance analysis, or model-specific interpretation. The pattern across these studies is not that one explanation technique is universally preferable, but that the explanation should match the information branch being inspected. Structured models can often be summarized globally with stable feature importance, whereas nonlinear language models require local approximation or token-level attribution that may vary across examples. This makes a branch-aware explanation strategy more defensible than applying one explanation method to every component of a heterogeneous system (De Magistris et al., 2022; Hashmi et al., 2024; Kozik et al., 2024; Muñoz & Iglesias, 2024; Nwaiwu et al., 2025).

Ensemble learning is relevant for the same reason. Different representations capture different parts of the misinformation problem, and stacking allows the errors of one base learner to be compensated by another without forcing every signal into a single feature space. Recent hybrid and multi-feature studies report that combining semantic and contextual information can improve classification when one source is ambiguous or incomplete. The benefit is especially plausible for LIAR, where claims are short and speaker history can be informative, although a final model should still show how much each source contributes to the decision (Aslam et al., 2025; Cao et al., 2025; Chalehchaleh et al., 2024; Farnoush et al., 2026; Koloski et al., 2022).

Metadata has a particular place in political fact-checking because prior statements and source-related indicators can act as compact signals of consistency. At the same time, such signals cannot replace the claim itself. A speaker with a poor previous record can still make a correct statement, while a generally reliable speaker can make a false one. The practical value of metadata therefore lies in combining contextual priors with current textual evidence and making disagreement visible when the two sources do not support the same outcome (Guess et al., 2020; Sharma & Singh, 2024; Zubiaga et al., 2018).

Evidence-oriented fact-checking research gives further support to this position. Automated verification becomes more useful when the final label is accompanied by information that allows a user to inspect why the claim was treated as questionable or credible. Argumentation and justification studies have similarly emphasized that verification should connect decisions with interpretable evidence rather than treat classification as an isolated output (Alhindi et al., 2018; Vallayil et al., 2023; X. Wang et al., 2025). In a text-metadata ensemble, disagreement between branches can therefore be treated as a diagnostic event that deserves attention rather than as a nuisance to be hidden by the final probability.

The broader misinformation field is also moving toward text-image and multimodal verification, where visual and linguistic evidence are combined in the same system. Those developments are important because online misinformation is not restricted to text, but they also increase explanation complexity. A smaller text-metadata setting remains useful for examining fusion behavior in a reproducible way before additional modalities are

introduced. The present implementation therefore keeps visual evidence outside the evaluated system and uses multimodal work only as a future extension rather than claiming a capability that was not tested (Jadhav et al., 2025; LekshmiAmmal & Madasamy, 2025; Lin et al., 2024; Moyo et al., 2026; Nasser et al., 2025).

Taken together, the literature leaves a practical gap between accurate hybrid detection and explanation of the fusion itself. High-performing models frequently report a final label and, in some cases, token or feature importance, but fewer studies inspect the relation between a semantic text branch and a structured source-history branch when their predictions diverge. This matters in journalistic verification because disagreement is precisely where a human reviewer may need more evidence, not less. The problem is therefore not only to raise classification performance, but to preserve a trace of how the two branches behave before and after fusion (Athira et al., 2023; Chalehchaleh et al., 2024; Muñoz & Iglesias, 2024; Nwaiwu et al., 2025; Vallayil et al., 2023; X. Wang et al., 2025).

Accordingly, the study aimed to determine whether combining text-based and metadata-based evidence improves binary veracity classification on LIAR while retaining explanations that show how the two branches contribute to, or disagree within, the final decision. Conceptually, the contribution is the treatment of disagreement between semantic claim evidence and structured speaker-history evidence as an interpretable signal for verification triage rather than as a simple classification error. The design links a BERT text branch, an Extreme Gradient Boosting metadata branch, an L1-regularized stacking layer, and branch-appropriate explanation methods in one evaluation, while keeping the text and metadata evidence separately inspectable.

Methods

A quantitative experimental design was used with the original LIAR benchmark partitions (see Table 1). The unit of analysis was one fact-checked political claim. The downloaded data contained 10,240 training, 1,284 validation, and 1,267 test observations, giving 12,791 claims in total. The six original truthfulness labels were remapped for binary classification as Fake=pants-fire, false, and barely-true, and Real=half-true, mostly-true, and true. The text branch used the claim statement, while the metadata branch used five speaker-history counts (barely-true, false, half-true, mostly-true, and pants-fire), party, state, job title, statement length, and three derived indicators: total claims, false rate, credibility score, and extreme rate. Missing categorical values were represented explicitly, job titles were restricted to the 100 most frequent categories with the remainder grouped as other, categorical encoders were fitted on the training partition, and the resulting 13 metadata features were standardized with training-fitted parameters. The BERT-base uncased tokenizer used a maximum sequence length of 128; the observed truncation rate was 0.03%.

The text classifier used BERT-base uncased with a 0.5:0.5 fusion of the [CLS] representation and mean-pooled token representation, multi-sample dropout, partial freezing of the lower encoder, AdamW optimization, focal loss, cosine learning-rate scheduling, and early stopping on validation macro-F1. The metadata classifier used Extreme Gradient Boosting, with hyperparameters selected by 60 Optuna trials under five-fold stratified cross-validation and the final estimator fitted on the training data with validation monitoring. For stacking, each branch supplied Fake and Real probabilities; absolute branch disagreement and the probability interaction were added as meta-features. Candidate meta-learners were compared on the validation predictions, and L1-regularized

logistic regression gave the selected test result. Performance was assessed on the untouched LIAR test partition with accuracy, macro-F1, class-wise F1, receiver operating characteristic area under the curve (ROC-AUC), average precision, and Matthews correlation coefficient (MCC), while paired errors of the ensemble and XGBoost were compared with McNemar's test. SHAP was used for the tree and stacking components, Local Interpretable Model-agnostic Explanations (LIME) for local text surrogates, and Integrated Gradients for token attribution; explanation assessment included SHAP deletion faithfulness, train-test SHAP stability, LIME surrogate R^2, Integrated Gradients convergence difference, and text-metadata disagreement analysis (Chen & Guestrin, 2016; Devlin et al., 2019; Lundberg & Lee, 2017; Ribeiro et al., 2016; Sundararajan et al., 2017).

Table 1. Summary of dataset, features, model components, and evaluation procedures

Component	Description
Dataset	LIAR benchmark; 10,240 train, 1,284 validation, 1,267 test claims.
Task	Binary classification: Fake vs Real.
Label mapping	Fake=pants-fire, false, barely-true; Real=half-true, mostly-true, true.
Text input	Claim statement; BERT-base uncased; maximum length 128 tokens.
Metadata input	Speaker-history counts, party, state, job title, statement length, and derived credibility variables.
Metadata preprocessing	Missing-category handling, top-100 job-title capping, training-fitted categorical encoding and standardization.
Text classifier	Fine-tuned BERT-base uncased.
Metadata classifier	Extreme Gradient Boosting; 60-trial. Optuna tuning with five-fold stratified cross-validation.
Stacking	Base probabilities +disagreement+interaction; selected L1-regularized logistic regression.
Explainability	SHAP, LIME, Integrated Gradients
Evaluation	Accuracy, macro-F1, class-wise F1, ROC-AUC, average precision, MCC, McNemar test.

Note: LIAR=benchmark dataset of fact-checked political statements; BERT=Bidirectional Encoder Representations from Transformers; SHAP=SHapley Additive exPlanations; LIME=Local Interpretable Model-agnostic Explanations; ROC-AUC=area under the receiver operating characteristic curve; MCC=Matthews correlation coefficient.

Figure 1 summarizes the implemented workflow. Claim text and structured metadata were processed in parallel rather than being collapsed into one input representation. The probability outputs were then combined at the stacking layer, while explanation methods were attached to the branch where they were most appropriate.

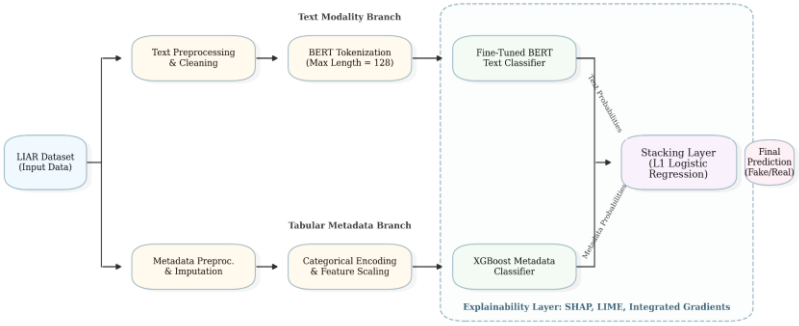

Figure 1. Overview of the explainable stacked ensemble using claim text and metadata (source: author's analysis).

Results

The LIAR data used in the experiments contained 12,791 short political claims across the original partitions: 10,240 training observations, 1,284 validation observations, and 1,267 held-out test observations. After binary remapping, 5,657 claims (44.2%) belonged to the Fake class and 7,134 (55.8%) to the Real class. Statements averaged 18.0 words (SD=10.1), and the maximum observed statement length was 467 words. With a BERT sequence limit of 128 tokens, only 0.03% of examples were truncated, so the chosen text length covered almost all claims without substantial loss of input content.

The six original labels were retained for diagnostic analysis even though the final target was binary. This made it possible to distinguish strongly false claims such as pants-fire from boundary categories such as barely-true and half-true after classification. A test example produced during the explanation stage was the statement, *"Wisconsin is on pace to double the number of layoffs this year,"* attributed in LIAR to Katrina Shankland. The ensemble assigned the Fake class with 0.9426 confidence, and both the text and metadata branches also favored Fake for this case. The example illustrates how the final prediction can be read together with branch-level evidence rather than as a label in isolation.

The text-only BERT model reached 0.6306 accuracy, 0.6203 macro-F1, 0.6547 ROC-AUC, and 0.2424 MCC on the held-out test set. The metadata-only XGBoost model was stronger, with 0.7277 accuracy, 0.7066 macro-F1, 0.8121 ROC-AUC, and 0.4458 MCC. The selected stacking model provided the highest overall balance, reaching 0.7514 accuracy, 0.7483 macro-F1, 0.7205 Fake-class F1, 0.7761 Real-class F1, 0.8288 ROC-AUC, 0.8640 average precision, and 0.4970 MCC. Relative to BERT, stacking improved macro-F1 by 0.1280 and ROC-AUC by 0.1741; relative to XGBoost, the increases were 0.0417 and 0.0167, respectively (see Table 2).

Table 2. Performance comparison on the LIAR test set for binary veracity classification

Model	Accuracy	F1 macro	F1 Fake	F1 Real	ROC-AUC	Avg. precision	MCC
BERT (text only)	0.6306	0.6203	0.5577	0.6829	0.6547	0.6957	0.2424
XGBoost (metadata)	0.7277	0.7066	0.6278	0.7853	0.8121	0.8522	0.4458

Ensemble (stacking)	0.7514	0.7483	0.7205	0.7761	0.8288	0.8640	0.4970

Note: F1 macro=macro-averaged F1-score; ROC-AUC=area under the receiver operating characteristic curve; Avg. precision=average precision score; MCC=Matthews correlation coefficient. Values are reported on the 1,267-observation held-out LIAR test partition.

The comparison against the strongest single branch was also examined with paired errors. McNemar's test for the ensemble and XGBoost gave a statistic of 95.0000 and p=0.0503. The numerical gains were therefore consistent across the reported aggregate metrics, but the paired error difference did not reach conventional statistical significance at α=0.05. The result is important for interpretation because it shows that the stacking gain was modest over an already strong metadata branch rather than a complete change in the error structure.

The selected L1 logistic stacking layer also made the fusion behavior inspectable. The largest positive coefficient belonged to the BERT-XGBoost probability interaction (+0.5666), followed by the complementary XGBoost class-probability coefficients (±0.5495). Absolute disagreement carried a positive coefficient of +0.2728, while the BERT probability pair had smaller coefficients of approximately ±0.1948. In the contribution analysis produced by the notebook, the metadata pathway accounted for 73.8% and the BERT pathway for 26.2%, which is consistent with the stronger standalone performance of XGBoost on LIAR.

Figure 2. Stacking performance diagnostics, including confusion matrix, discrimination curves, confidence behavior, and calibration (source: author's analysis)

Figure 2 shows that the ensemble improvement was not restricted to one summary score. The confusion matrix contained 406 correctly classified Fake claims and 546 correctly classified Real claims, with 147 Fake claims classified as Real and 168 Real

claims classified as Fake. The ROC and precision-recall views place the ensemble above the text-only branch across much of the operating range, while calibration and confidence plots show how the three models distribute probability rather than only hard labels. These diagnostics support the interpretation of the ensemble as a balancing layer over two unequally strong sources of evidence.

The metadata branch was the clearest source of predictive strength. SHAP ranking placed credibility_score first with mean |SHAP|=0.6929, followed by barely_true_ct=0.3290, false_rate=0.3006, total_claims=0.1526, half_true_ct=0.1408, false_ct=0.1154, statement words=0.0876, and extreme_rate=0.0559. These variables are mainly speaker-history or history-derived indicators, showing that prior fact-check patterns were strongly associated with the remapped LIAR labels. Figure 3 provides the correlation structure and class-wise false-rate distribution used to inspect how those variables relate to one another.

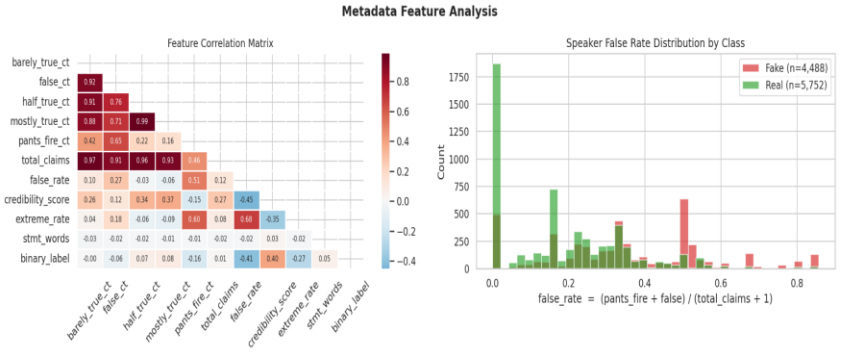

Figure 3. Correlation structure and class-wise distribution of key metadata variables (source: author's analysis).

The explanation assessment produced different levels of evidence for the structured and textual components. In the XGBoost deletion test, the baseline ROC-AUC of 0.8121 fell to 0.5000 after features were removed in SHAP importance order, giving an overall degradation of 0.3121 and a deletion-curve area of 0.5883. Train-test SHAP importance rankings were highly stable, with Pearson ρ=0.9991. The local LIME surrogates produced R^2 values of 0.4413, 0.5525, 0.2435, and 0.1942, for a mean of 0.3579 (SD=0.1455), which indicates moderate rather than near-perfect local fidelity. Integrated Gradients convergence differences were 0.1657, 1.1035, 0.3720, and 0.5804, with mean |δ|=0.5554, so those token-level attributions were used as local diagnostics rather than as a claim of complete explanation fidelity (see Table 3).

Table 3. Explainability and disagreement diagnostics

Method/category	Metric	Score
SHAP TreeExplainer.	Faithfulness (AUC degradation).	0.3121
SHAP TreeExplainer	Stability (train vs. test ρ)	0.9991
LIME.	Local fidelity (surrogate R^2).	0.3579
Integrated Gradients.	Mean convergence |δ|	0.5554
Text-metadata disagreement.	High-disagreement cases.	385 (30.4%)
Text-metadata disagreement.	XGBoost accuracy on subset.	0.7922
Text-metadata disagreement.	BERT accuracy on subset.	0.4649

Text-metadata disagreement.	Ensemble accuracy on subset.	0.8234

Note: AUC degradation=loss of discrimination when features are deleted in SHAP importance order; ρ=Pearson correlation between train and test global SHAP importance; surrogate R^2=local fit of the LIME surrogate; |δ|=absolute Integrated Gradients convergence difference. High disagreement was defined as an absolute difference greater than 0.3 between the BERT and XGBoost Real-class probabilities.

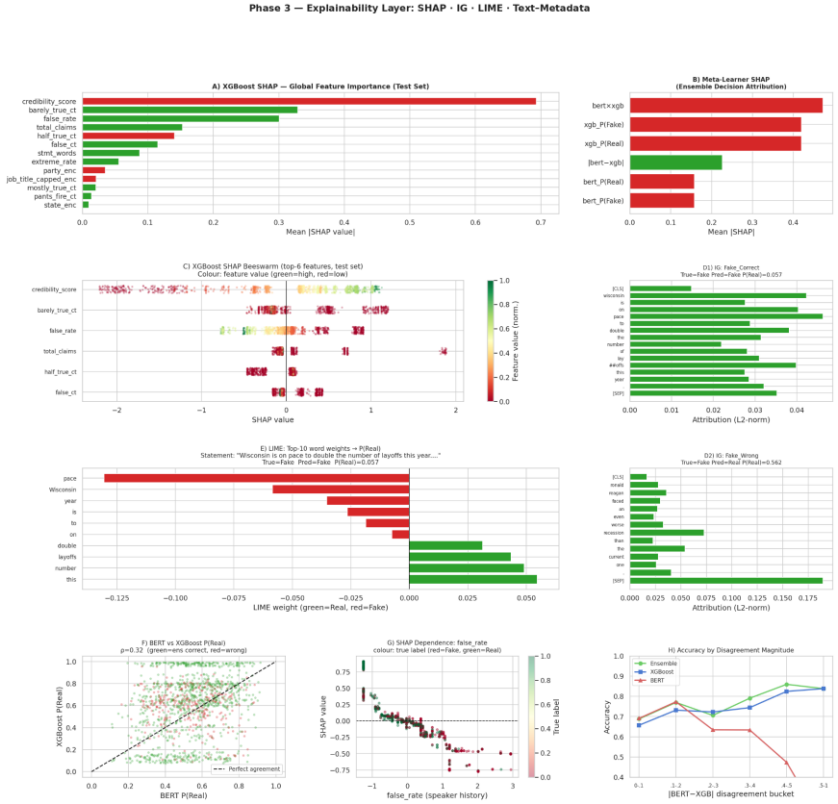

Figure 4. Explainability diagnostics and text-metadata disagreement behavior (source: author's analysis)

Figure 4 shows text-metadata disagreement supplied an additional view of the system that could not be obtained from overall accuracy alone. Using |P_BERT(Real) - P_XGBoost(Real)| > 0.3 as the high-disagreement rule, 385 test claims, or 30.4% of the test set, met the criterion. On these cases, BERT accuracy was 0.4649, XGBoost accuracy was 0.7922, and ensemble accuracy was 0.8234. XGBoost was correct while BERT was wrong in 202 cases, whereas BERT was correct while XGBoost was wrong in 76 cases. On the lower-disagreement cases, the corresponding accuracies were 0.7029 for BERT, 0.6995 for XGBoost, and 0.7200 for the ensemble. The high-disagreement subset therefore represented a specific zone where the two information sources behaved differently and where fusion was most useful.

Error rates also differed substantially across the six original truthfulness categories. Pants-fire claims had the lowest error rate at 0.098, followed by mostly-true at 0.174. False, half-true, and true produced intermediate rates of 0.241, 0.264, and 0.269, whereas barely-true was the most difficult category at 0.368. This pattern is compatible with the binary remapping: boundary labels carry more semantic ambiguity than the strongest false category. The notebook also identified 63 high-confidence errors, showing that high

posterior probability did not guarantee correctness and reinforcing the need to keep explanation and confidence information visible in a verification workflow (see Table 4).

Table 4. Error rate by original LIAR veracity label

Veracity label	Error rate	N
pants-fire	0.098	92
false	0.241	249
barely-true	0.368	212
half-true	0.264	265
mostly-true	0.174	241
true	0.269	208

Note: Error rate=fraction of misclassified samples within each original LIAR veracity category; n=number of test observations in that category. The category counts sum to the 1,267-observation test partition.

Across the results, three observations were consistent. The stacking architecture produced the strongest aggregate classification performance, metadata supplied most of the predictive strength in the current LIAR setting, and explanation diagnostics exposed where the two sources supported or contradicted one another. The disagreement analysis was especially informative because it separated ordinary cases from the 30.4% of test observations where the branch probabilities diverged sharply. In those cases, the ensemble did not simply average the branches; its learned interaction and disagreement terms allowed the final model to use the structure of the conflict itself.

Discussion

The central result is that the text-metadata combination improved the balance of binary veracity classification without removing the distinction between the two information sources. The ensemble macro-F1 of 0.7483 was higher than both the BERT branch (0.6203) and XGBoost branch (0.7066), while ROC-AUC increased to 0.8288. The gain over XGBoost was numerically smaller than the gain over BERT, and the McNemar result of p=0.0503 did not cross the conventional 0.05 threshold. The interpretation should therefore rest on the full pattern rather than on a claim of decisive statistical superiority. Stacking improved class balance, average discrimination, and MCC, but it did so on top of a metadata model that was already strong on this benchmark.

The strength of metadata is one of the most important findings for understanding LIAR. The claims are short political statements, so the language available to BERT is often limited, ambiguous, or dependent on facts outside the sentence. Speaker-history counts and derived credibility variables add a different kind of information: they summarize patterns associated with who made previous statements and how those statements had been rated. This explains why credibility_score, barely_true_ct, and false_rate were more influential than simple statement length. The result is consistent with work showing that source and contextual information can complement content-based misinformation detection (Chalehchaleh et al., 2024; Koloski et al., 2022; Sharma & Singh, 2024).

At the same time, stronger metadata does not mean that the text branch is unnecessary. The disagreement analysis shows why. There were 76 test cases where BERT was correct and XGBoost was wrong, so contextual history did not always provide the better answer. Conversely, XGBoost corrected many more BERT failures, which explains the larger metadata contribution in the learned stack. A verification system that exposes both scores can therefore distinguish between agreement, metadata-led

disagreement, and text-led disagreement. This is more useful for editorial triage than a single probability because it shows when the decision rests mainly on historical context and when the current wording resists that context.

The 385 high-disagreement cases are particularly relevant to journalistic use. Their ensemble accuracy of 0.8234 exceeded the two individual branches, and the difference between branch probabilities gives an immediate signal that the evidence is not internally uniform. Such cases can be prioritized for human review because the model is effectively reporting that the current claim and the speaker-history context point in different directions. This use of disagreement does not turn the classifier into an automated fact-checker with external evidence; instead, it provides a practical cue for where a journalist may need to verify sources, search supporting documents, or examine the speaker's prior record before accepting the prediction.

The explanation results also show why explanation quality should not be treated as one number. TreeSHAP behaved strongly for structured metadata: deletion caused a 0.3121 loss in ROC-AUC and global importance rankings were almost unchanged between train and test (ρ=0.9991). LIME behaved differently. Its mean surrogate R^2 of 0.3579 indicates moderate local fidelity, which is reasonable for a nonlinear transformer but does not justify presenting each local surrogate as an exact account of BERT reasoning. Integrated Gradients supplied token-level direction, while its mean convergence difference of 0.5554 advises similar caution. The practical interpretation is therefore layered: SHAP supports robust global inspection of metadata and fusion, while LIME and Integrated Gradients provide local textual clues that should be read as diagnostics rather than ground truth about semantic reasoning (Athira et al., 2023; Guidotti et al., 2018; Vallayil et al., 2023).

This layered behavior addresses an interpretability issue that is often hidden in hybrid detection studies. A single salience plot cannot explain a system whose branches use fundamentally different data. The metadata model asks which contextual variables moved the probability, the text model asks which parts of a claim influenced the local prediction, and the stack asks how those two probabilities interacted. Keeping these questions separate avoids the impression that one explanation method has captured the whole decision process. It also makes the final model easier to audit because an unusual prediction can be traced to a source-history feature, a textual cue, an unusually large branch disagreement, or a strong interaction term.

The error distribution across the original LIAR labels adds another explanation for the observed performance. Barely-true claims were the hardest category after remapping, with an error rate of 0.368, while pants-fire claims were much easier at 0.098. Binary classification creates a sharp boundary around labels that were originally ordinal and nuanced. A barely-true statement can contain a mixture of correct and incorrect information, while pants-fire is a stronger false category, so the former is naturally harder to place once only Fake and Real remain. This pattern also helps explain why a model that uses both text and historical context can improve macro-F1 without eliminating uncertainty near the original label boundaries.

For journalism, the useful contribution is not the automation of editorial judgment. The value lies in separating routine cases from cases that deserve attention and in providing interpretable reasons for that separation. Prior studies of AI in journalism emphasize that professionals expect transparency, human agency, and accountable use of automated assistance rather than opaque replacement of editorial decisions (Broussard et al., 2019; Gutiérrez-Caneda et al., 2024; Luttrell et al., 2025; Palla & Kostarella, 2025).

A text-metadata disagreement flag fits that role because it can be used as a screening signal: agreement may support routine triage, while large disagreement can prompt additional checking before any editorial conclusion is made.

The findings also place the study between text-only classification and broader multimodal misinformation systems. Current multimodal work combines text with images and other media and can capture evidence that is absent from LIAR. That direction is valuable, but it creates additional questions about how cross-source reasoning should be explained to users (Jadhav et al., 2025; LekshmiAmmal & Madasamy, 2025; Lin et al., 2024; Nasser et al., 2025). The present text-metadata setting isolates one fusion problem and provides a measurable disagreement signal. That signal can later be extended to additional evidence types, but such an extension should be evaluated rather than assumed from the current results.

Several boundaries should be kept in view when interpreting the results. LIAR is one benchmark composed of short United States political claims, so the reported values should not be generalized directly to full-length news articles, other countries, or multimodal misinformation without new evaluation. The binary remapping also compresses six truthfulness categories into two classes, which reduces the nuance of boundary labels such as barely-true and half-true. In addition, speaker-history metadata is most informative when comparable contextual history exists; performance may change for new or sparsely represented speakers. Finally, LIME and Integrated Gradients were examined on representative local cases and no newsroom user study was conducted, so the operational usefulness of those explanations should be tested with journalists in later work. These points define the scope of the evidence while leaving the reported LIAR experiment and its observed text-metadata behavior unchanged.

Within those boundaries, the results answer the original objective in two parts. Combining the branches increased macro-F1 and ROC-AUC compared with either standalone input pathway, while explanation diagnostics retained visibility into the contribution and conflict of those pathways. The most distinctive evidence came from high disagreement, where the stack reached 0.8234 accuracy and outperformed both base learners. This indicates that disagreement can be used as more than an error description: it can become a practical indicator of cases in which content and context deserve separate inspection.

Conclusion

Combining claim text with structured speaker metadata improved binary veracity classification on the LIAR test set while retaining separate evidence about how the two branches behaved. The selected L1-regularized stacking model reached 0.7514 accuracy, 0.7483 macro-F1, 0.8288 ROC-AUC, and 0.4970 MCC, compared with macro-F1 values of 0.6203 for BERT and 0.7066 for XGBoost. Metadata supplied the larger share of predictive strength, particularly through credibility and speaker-history indicators, but the text branch still corrected a subset of metadata errors. The most useful diagnostic result was the 30.4% high-disagreement subset, where ensemble accuracy reached 0.8234 and the difference between semantic and contextual evidence became directly visible. For journalistic verification, the recommended use is therefore decision support rather than automatic editorial judgment: large branch disagreement can be treated as a cue for additional source checking, while SHAP, LIME, and Integrated Gradients provide complementary levels of inspection. Future evaluation should test the same disagreement principle on additional misinformation datasets, new or low-history speakers, and later

text-image settings, while preserving a clear distinction between the evidence contributed by each branch.

Authors' Contributions and Responsibilities

The authors made substantial contributions to the conception and design of the study. The authors took responsibility for data analysis, interpretation, and discussion of results. The authors read and approved the final manuscript.

Conflict of Interest

The authors certify that there is no conflict of interest arising from financial, personal, or other relationships relevant to the material discussed in the manuscript.

Generative AI Statement

Grammarly was used during manuscript preparation as an AI-assisted language-editing tool solely for language polishing, including improvements in grammar, clarity, and readability. All suggested changes were critically reviewed and edited by the authors, who take full responsibility for the scientific content, analysis, interpretation, citations, and final manuscript.

Acknowledgements

The authors gratefully acknowledge the encouragement and support of their families during the preparation of this article.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alghamdi, J., Lin, Y., & Luo, S. (2024). Unveiling the hidden patterns: A novel semantic deep learning approach to fake news detection on social media. *Engineering Applications of Artificial Intelligence*, *137*, 109240. https://doi.org/10.1016/j.engappai.2024.109240

Alhindi, T., Petridis, S., & Muresan, S. (2018). Where is Your Evidence: Improving Fact-checking by Justification Modeling. In J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, & A. Mittal (Eds.), *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (pp. 85–90). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5513

Aslam, Z., Missen, M. M. S., Ghaffar, A. A., Mehmood, A., Villar, M. G., Alvarado, E. S., & Ashraf, I. (2025). Advancing fake news combating using machine learning: A hybrid model approach. *Knowledge and Information Systems*, *67*(12), 12137–12177. https://doi.org/10.1007/s10115-025-02588-y

Athira, A. B., Kumar, S. D. M., & Chacko, A. M. (2023). A systematic survey on explainable AI applied to fake news detection. *Engineering Applications of Artificial Intelligence*, *122*, 106087. https://doi.org/10.1016/j.engappai.2023.106087

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., & Chuan, C.-H. (2019). Artificial Intelligence and Journalism. *Journalism & Mass Communication Quarterly*, *96*(3), 673–695. https://doi.org/10.1177/1077699019859901

Calvo-Rubio, L.-M., & Ufarte-Ruiz, M.-J. (2021). Artificial intelligence and journalism: Systematic review of scientific production in Web of Science and Scopus (2008-2019). *Communication & Society*, 159–176. https://doi.org/10.15581/003.34.2.159-176

Cao, J., Zhuo, S., Su, J., & Chen, G. (2025). A Fake News Detection Model Based on Capsule Networks and Collaborative Attention. *Applied Sciences*, *15*(22), 12190. https://doi.org/10.3390/app152212190

Chalehchaleh, R., Salehi, M., Farahbakhsh, R., & Crespi, N. (2024). BRaG: A hybrid multi-feature framework for fake news detection on social media. *Social Network Analysis and Mining*, *14*(1), 35. https://doi.org/10.1007/s13278-023-01185-7

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, 785–794. https://doi.org/10.1145/2939672.2939785

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2021). Trends in combating fake news on social media – a survey. *Journal of Information and Telecommunication*, *5*(2), 247–266. https://doi.org/10.1080/24751839.2020.1847379

De Magistris, G., Russo, S., Roma, P., Starczewski, J. T., & Napoli, C. (2022). An Explainable Fake News Detector Based on Named Entity Recognition and Stance Classification Applied to COVID-19. *Information*, *13*(3), 137. https://doi.org/10.3390/info13030137

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Farnoush, A., Gupta, A., Li, H., Benitez, J., & Jiang, W. (Kayla). (2026). Towards developing fake and satire news detection policies using component-based SEM and interpersonal detection theory. *Information & Management*, *63*(2), 104293. https://doi.org/10.1016/j.im.2025.104293

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

Gutiérrez-Caneda, B., Lindén, C.-G., & Vázquez-Herrero, J. (2024). Ethics and journalistic challenges in the age of artificial intelligence: Talking with professionals and experts. *Frontiers in Communication*, *9*. https://doi.org/10.3389/fcomm.2024.1465178

Hashmi, E., Yayilgan, S. Y., Yamin, M. M., Ali, S., & Abomhara, M. (2024). Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI. *IEEE Access*, *12*, 44462–44480. https://doi.org/10.1109/ACCESS.2024.3381038

Hu, B., Mao, Z., & Zhang, Y. (2025). An overview of fake news detection: From a new perspective. *Fundamental Research*, *5*(1), 332–346. https://doi.org/10.1016/j.fmre.2024.01.017

Jadhav, R., Meshram, V., Bhosle, A., Patil, K., Dash, S., & Jadhav, S. (2025). Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1690616

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

Koloski, B., Stepišnik Perdih, T., Robnik-Šikonja, M., Pollak, S., & Škrlj, B. (2022). Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*, *496*, 208–226. https://doi.org/10.1016/j.neucom.2022.01.096

Kozik, R., Ficco, M., Pawlicka, A., Pawlicki, M., Palmieri, F., & Choraś, M. (2024). When explainability turns into a threat—Using xAI to fool a fake news detection method. *Computers & Security*, *137*, 103599. https://doi.org/10.1016/j.cose.2023.103599

Kula, S., Kozik, R., & Choraś, M. (2022). Implementation of the BERT-derived architectures to tackle disinformation challenges. *Neural Computing and Applications*, *34*(23), 20449–20461. https://doi.org/10.1007/s00521-021-06276-0

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

LekshmiAmmal, H. R., & Madasamy, A. K. (2025). A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. *Journal of Big Data*, *12*(1), 46. https://doi.org/10.1186/s40537-025-01093-x

Lin, S.-Y., Chen, Y.-C., Chang, Y.-H., Lo, S.-H., & Chao, K.-M. (2024). Text–image multimodal fusion model for enhanced fake news detection. *Science Progress*, *107*(4), 00368504241292685. https://doi.org/10.1177/00368504241292685

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874

Luttrell, R., Davis, J., & Welch, C. (2025). Source attribution and detection strategies for AI-era journalism. *Telecommunications Policy*, *49*(10), 103053. https://doi.org/10.1016/j.telpol.2025.103053

Mishima, K., & Yamana, H. (2022). A Survey on Explainable Fake News Detection. *IEICE Transactions on Information and Systems*, *E105.D*(7), 1249–1257. https://doi.org/10.1587/transinf.2021EDR0003

Moyo, B. V., Tuyikeze, T., Matsebula, F., & Obagbuwa, I. C. (2026). An AI-driven conceptual framework for detecting fake news and deepfake content: A systematic review. *Frontiers in Artificial Intelligence*, *9*. https://doi.org/10.3389/frai.2026.1737790

Muñoz, S., & Iglesias, C. Á. (2024). Exploiting Content Characteristics for Explainable Detection of Fake News. *Big Data and Cognitive Computing*, *8*(10), 129. https://doi.org/10.3390/bdcc8100129

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007. https://doi.org/10.1016/j.jjimei.2020.100007

Nasser, M., Arshad, N. I., Ali, A., Alhussian, H., Saeed, F., Da'u, A., & Nafea, I. (2025). A systematic review of multimodal fake news detection on social media using deep learning models. *Results in Engineering*, *26*, 104752. https://doi.org/10.1016/j.rineng.2025.104752

Nwaiwu, S., Jongsawat, N., & Tungkasthan, A. (2025). Decoding Disinformation: A Feature-Driven Explainable AI Approach to Multi-Domain Fake News Detection. *Applied Sciences*, *15*(17), 9498. https://doi.org/10.3390/app15179498

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, *1*(3). https://doi.org/10.37016/mr-2020-024

Palla, Z., & Kostarella, I. (2025). Journalists' Perspectives on the Role of Artificial Intelligence in Enhancing Quality Journalism in Greek Local Media. *Societies*, *15*(4), 89. https://doi.org/10.3390/soc15040089

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Rai, N., Kumar, D., Kaushik, N., Raj, C., & Ali, A. (2022). Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering*, *3*, 98–105. https://doi.org/10.1016/j.ijcce.2022.03.003

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sharma, U., & Singh, J. (2024). A comprehensive overview of fake news detection on social networks. *Social Network Analysis and Mining*, *14*(1), 120. https://doi.org/10.1007/s13278-024-01280-3

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, *8*(3), 171–188. https://doi.org/10.1089/big.2020.0062

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

Tandoc Jr., E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Vallayil, M., Nand, P., Yan, W. Q., & Allende-Cid, H. (2023). Explainability of Automated Fact Verification Systems: A Comprehensive Review. *Applied Sciences*, *13*(23), 12608. https://doi.org/10.3390/app132312608

Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short*

Papers) (pp. 422–426). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2067

Wang, X., Cabrio, E., & Villata, S. (2025). When automated fact-checking meets argumentation: Unveiling fake news through argumentative evidence. *Argument & Computation*, *16*(3), 405–424. https://doi.org/10.1177/19462174251330980

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 109:1-109:40. https://doi.org/10.1145/3395046

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys (CSUR)*, *51*(2), 32:1-32:36. https://doi.org/10.1145/3161603