A Unified Explainable Fake News Detection System based on Transformer-Driven Natural Language Understanding

Mostafa Farzana
Nanjing University of Information Science and Technology
Nanjing, China

Iftee Shekh Iftesham
Nanjing University of Information Science and Technology
Nanjing, China

Tanha Asfika Jaman
Nanjing University of Information Science and Technology
Nanjing, China

Akter Mst Lovely
Nanjing University of Information Science and Technology
Nanjing, China

ABSTRACT
Online misinformation calls for classifiers that are both accurate and open to inspection. This paper presents a text-only fake news detection system that combines WordPiece tokenization, a fine-tuned BERT encoder, a linear softmax classification head, and Local Interpretable Model-Agnostic Explanations (LIME). The model is trained on a merged corpus of two public files of full-length articles, 52.3% fake and 47.7% real, whose composition is consistent with the ISOT Fake News Dataset. On a held-out test subset of 500 articles (245 fake, 255 real), the classifier makes no errors: precision, recall, F1, accuracy, the Matthews correlation coefficient, and ROC-AUC all equal 1.00. Because a perfect score on a finite sample is not proof of a zero error rate, the results are analyzed further with exact 95% confidence intervals, a zero-error sensitivity analysis, a class-proportion check of the test subset, and a comparison with results reported on other benchmark datasets. The analysis shows that, under a binomial sampling assumption, the in-domain error rate is below 0.6% with 95% confidence, that no statistically significant difference between the class proportions of the test subset and the corpus was detected ($p = 0.14$), and that ceiling-level scores are typical of binary article corpora rather than of harder benchmarks. LIME explanations make individual predictions inspectable and can expose source-related cues. The system is therefore a transparent baseline for explainable news-text classification, and the paper defines the cross-dataset, source-disjoint, and temporal evaluations required before deployment.

General Terms
Algorithms, Experimentation, Reliability

Keywords
Fake news detection, BERT, transformer, explainable artificial intelligence, LIME, natural language understanding

1. INTRODUCTION
False and misleading news can spread rapidly through online platforms, where manual verification cannot keep pace with the volume of content [1]. Automated detection is therefore useful as a screening aid, but the task is difficult: fabricated and legitimate articles may share vocabulary and journalistic structure, and a text classifier can learn topic, source, or style correlations instead of credibility. Research on misinformation has consequently moved from sparse lexical features toward contextual neural representations and richer benchmark resources [2], [3], [4].

This work frames fake news detection as supervised binary text classification and combines a pretrained BERT encoder with a lightweight classification head [5]. The model predicts a fake or real label from article text, while LIME approximates each individual decision with locally weighted token contributions [6]. A classifier of this kind does not verify events against external evidence; it learns statistical regularities in labeled text. Contextual representations improve sensitivity to wording and long-range dependencies, but they can also encode publication identity, topic distribution, names, and other shortcuts [7]. An explainable interface is valuable precisely because it can reveal such dependencies and allow a reviewer to question them.

The contributions of this paper are fourfold. First, it defines a compact end-to-end architecture from article input to probability output and local explanation. Second, it reports the held-out performance of this architecture on a merged fake/real news-article corpus with a full set of threshold-based and ranking metrics. Third, it analyzes the statistical reliability of a perfect test score through exact confidence intervals, a zero-error sensitivity analysis, and a class-proportion check of the test subset. Fourth, it places the result in a cross-dataset context and uses local token attribution to identify likely dataset artifacts, which together explain why ceiling-level scores on one corpus do not imply general misinformation detection.

The remainder of the paper reviews related methods, defines the task, describes the dataset and preprocessing, presents the BERT-LIME pipeline, details the experimental setup and evaluation scenarios, analyzes the results and explanations, and states the limitations that govern responsible use.

2. RELATED WORK
Early fake news classifiers used bag-of-words, n-grams, TF-IDF, sentiment, and readability features with logistic regression, support vector machines, decision trees, or related algorithms. These models are efficient but depend strongly on feature design and can degrade when topics and sources change [2], [3]. Neural document classifiers reduced manual feature engineering: convolutional models captured local patterns, while recurrent and hierarchical attention networks modeled sequences and document structure [8]. Their sequential computation and limited transparency remained practical concerns.

Transformers replaced recurrence with self-attention, enabling

parallel processing and direct interaction among distant tokens [9]. BERT pretraining produces bidirectional contextual representations that can be adapted to classification with a task-specific head [5]. BERT-based fake news systems have reported strong benchmark performance [10], while repositories such as FakeNewsNet emphasize that news content, social context, and temporal signals can all matter [4]. The present system deliberately remains text-only, so its claims are limited to linguistic evidence.

Post-hoc explanation methods address the opacity of high-capacity classifiers. LIME perturbs an input and fits a sparse local surrogate around the prediction [6]; SHAP provides an alternative additive attribution framework [11]. Local explanations can expose influential tokens, but they do not establish factual reasoning or global reliability. The gap addressed here is therefore operational: the system joins a BERT classifier, probability output, local token attribution, and an interface in one workflow, and evaluates that workflow together with an explicit analysis of what its test evidence can and cannot support.

2.1 Method Positioning

The main methodological families differ in what they represent and what a reviewer can inspect. Sparse lexical models offer transparent feature weights but are sensitive to vocabulary and topic changes. Neural sequence and document models learn representations with less manual engineering, while transformers model long-range context efficiently. Post-hoc attribution adds instance-level inspection without converting a correlational classifier into a fact-verification system. Table 1 positions the proposed workflow within these tradeoffs; it is a conceptual comparison grounded in the cited literature.

Table 1. Positioning of Text-Classification Approaches

Approach	Main capability	Important boundary
Sparse lexical [2], [3]	Efficient, explicit word-feature models	Sensitive to topic and vocabulary shift
CNN / RNN / HAN [8]	Learns local, sequential, or hierarchical patterns	Limited transparency; recurrent computation
BERT [5], [10]	Bidirectional contextual representation	May learn source or topic shortcuts
BERT + LIME (this work)	Contextual prediction with local token audit	Explanation is local and post hoc

The pipeline integrates components that are individually established: BERT supplies contextual encoding, a linear head supplies binary decisions, and LIME supplies a local surrogate. The contribution is their consolidated and inspectable workflow and its careful evaluation, not a new transformer architecture or a new explanation algorithm.

2.2 Dataset Dependence of Reported Performance

Fake news resources differ substantially in label definition and granularity. LIAR contains about 12.8K short political statements with six fine-grained truthfulness labels [3]; FakeNewsNet combines news content with social context and spatiotemporal information [4]; and ISOT contains full articles whose real class is drawn from a single wire service [12], [13]. A model trained on one construction may solve a different statistical problem from a model trained on another. Table 2 summarizes results reported in the literature on these resources. The figures are not directly comparable because tasks, label sets, and splits differ, but their spread shows that the dataset, rather than the model alone, largely determines the attainable score.

Table 2. Reported Results Across Fake News Benchmarks

Dataset	Task	Method	Reported result
LIAR [3]	Short statements, 6 classes	Hybrid CNN	27.4% accuracy
ISOT (early version) [12]	Articles, binary	TF-IDF + linear SVM	92% accuracy
Kaggle fake news [10]	Articles, binary	FakeBERT	98.90% accuracy
Merged article corpus (this work)	Articles, binary	BERT + LIME	100% accuracy (n = 500)

The contrast is instructive. Statement-level, six-way classification on LIAR remains far from solved, whereas binary article-level classification on ISOT and similar Kaggle collections routinely approaches the ceiling. A perfect score on such a corpus therefore says more about its separability than about general misinformation detection, which motivates the reliability and artifact analyses in Sections 7 and 8.

3. PROBLEM FORMULATION

Let each article be a text x_i with binary label y_i, where 0 denotes fake and 1 denotes real. The labeled corpus is

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N},\ y_i \in \{0,1\} \tag{1}$$

where N is the number of labeled articles. A parameterized encoder and classifier produce one logit per class, $z = (z_0, z_1)$. Softmax converts them to posterior probabilities for each class $c \in \{0, 1\}$:

$$p_\theta(y = c \mid x) = \frac{e^{z_c}}{\sum_{k=0}^{1} e^{z_k}},\ c \in \{0,1\} \tag{2}$$

where z_c is the logit of class c and the denominator sums over both classes, k = 0 and k = 1. The predicted label is the class with the larger posterior. Training minimizes categorical cross-entropy, where y_{ic} is the one-hot target and p_{ic} is the predicted probability for article i and class c:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=0}^{1} y_{ic} \log p_{ic} \tag{3}$$

This formulation yields class probabilities for evaluation and for LIME queries. The probabilities are model scores, not verified truth estimates; calibration requires a separate reliability analysis [14].

Each document is converted to a token sequence that begins with [CLS] and uses [SEP] to end or separate segments. Shorter sequences are padded and longer sequences are truncated at a maximum length. If t_j is a valid token, its attention-mask entry is 1; padding positions receive 0. Because news articles are often longer than the encoder window, truncation means that the classifier sees mainly the opening of each article, which is also where datelines and source attributions appear.

4. DATASET AND PREPROCESSING

4.1 Corpus Description

The experiments use two publicly available collections of full article text, distributed as one file of fake articles and one file of real articles. Each file was reduced to article text and label, the two files were concatenated and shuffled, and the result was split into training, validation, and test subsets. Labels follow 0 = fake and 1 = real, and the merged corpus is 52.3% fake and 47.7% real. This two-file structure and these class proportions coincide with the published composition of the ISOT Fake News Dataset [12], [13], which contains 23,481 fake and 21,417 real articles

[15] and whose real articles were collected from Reuters.com, while fake articles came from sites flagged as unreliable by PolitiFact. The identification rests on these characteristics, and statements below that depend on it are marked as conditional. Table 3 and Figure 1 summarize the corpus.

The small difference between class shares (4.6 percentage points) removes any meaningful majority-class advantage: a classifier that always predicts fake would reach only 52.3% accuracy. Class balance does not, however, control topic, source, time, or writing-style differences. If the real class is drawn from a single publisher, as in ISOT, house-style features such as the "(Reuters)" dateline, location prefixes, and consistent formatting would be perfectly correlated with the real label. Such features are legitimate signals within the corpus but would not transfer to real news from other outlets.

Table 3. Corpus Summary

Property	Value
Source files	One fake-article file, one real-article file
Fake share (label 0)	52.3%
Real share (label 1)	47.7%
Model input	Article text only
Held-out test subset	500 articles (245 fake, 255 real)
Published ISOT composition, for comparison [15]	23,481 fake / 21,417 real (52.3% / 47.7%)

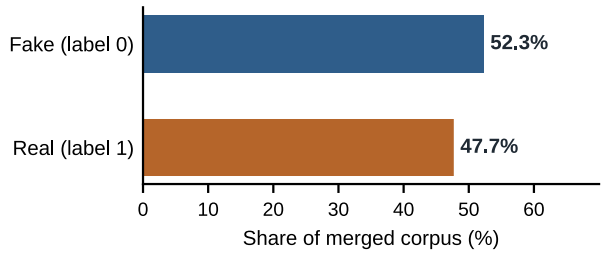


Figure 1. Class proportions of the merged fake and real article corpus.

4.2 Preprocessing and Tokenization

Preprocessing is intentionally minimal: Unicode normalization, removal of non-textual artifacts, and length control through truncation and padding. Stop-word removal, stemming, and lemmatization are not applied because they conflict with the minimal-preprocessing design and can remove cues that BERT learned during pretraining. WordPiece tokenization maps text to subword identifiers and inserts the classification and separator tokens used by BERT, and attention masks distinguish valid tokens from padding. Minimal processing also keeps the input consistent with the pretrained tokenizer: case, punctuation, morphology, and rare word fragments may carry information, and WordPiece can represent unseen words as known subword units. The same normalization and tokenizer must be applied at inference, since any additional cleaning would change the input distribution.

4.3 Split Design and Leakage Considerations

Merging only the text and the target label prevents direct leakage from explicit metadata fields such as subject and date, which can be class-specific. It does not remove information embedded in the prose, such as wire-service names, locations, bylines, recurring topics, or formatting patterns. The two files may also contain duplicated or closely related reports, and a random article-level split can place such items on both sides of the partition.

A leakage-resistant protocol would retain a stable identifier and a hash of each raw article before normalization, so that exact and near-duplicates can be grouped into one subset. A stronger evaluation would additionally separate publishers or time periods so that the test set measures transfer rather than memorization of source signatures. These scenarios are defined in Section 6.3, and the class proportions of the random test subset used here are checked in Section 7.3.

5. PROPOSED METHOD

5.1 BERT Classification Pipeline

The system passes token identifiers and attention masks through a pretrained BERT encoder. Self-attention forms contextual representations by projecting the input matrix X into queries, keys, and values. For key dimension d_k,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q, K, and V are the query, key, and value matrices obtained by learned linear projections of X, and division by $\sqrt{d_k}$ keeps the dot products in a numerically stable range. Multi-head attention repeats this operation in parallel representation subspaces, and residual connections, layer normalization, and position-wise feed-forward layers refine each token representation [9]. The final hidden state of the [CLS] token, $h_{[CLS]}$, serves as the article representation, and a linear head produces the two logits:

$$z = Wh_{[CLS]} + b \tag{5}$$

where $h_{[CLS]} \in \mathbb{R}^d$ is the final [CLS] hidden state of dimension d, $W \in \mathbb{R}^{2\times d}$ and $b \in \mathbb{R}^2$ are the weight matrix and bias of the classification head, and $z = (z_0, z_1)$ are the logits used in Equation (2). All encoder layers and the classification head are fine-tuned jointly. At inference, the same tokenizer and masking procedure are applied, the softmax scores are returned, and the selected article can be passed to the explanation module. An encoder-only design is appropriate because the task requires representation learning rather than text generation: bidirectional self-attention lets each retained token use both left and right context, and attention shortens the computational path between distant tokens compared with recurrent networks. Figure 2 shows the classification and explanation paths.

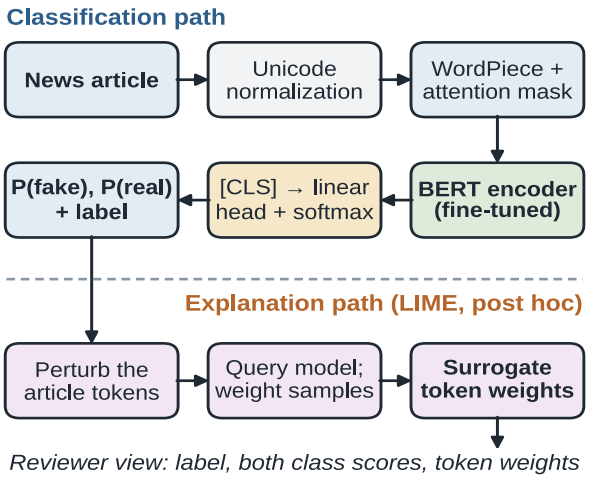


Figure 2. Architecture of the explainable BERT classification and LIME explanation workflow.

5.2 Local Explanation with LIME

For an article x, LIME creates perturbed token-presence vectors, queries the BERT classifier, weights the perturbations by their proximity to x, and fits an interpretable surrogate g. The local explanation is selected by

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g) \qquad (6)$$

Here f is the classifier, G is the family of interpretable surrogate models (sparse linear models over token presence), L measures the local lack of fidelity of g to f under the locality kernel π_x, and $\Omega(g)$ penalizes surrogate complexity [6]. For text, removing or masking selected tokens creates a neighborhood around the original article, the classifier's probabilities become the targets of the surrogate, and proximity weighting gives more influence to perturbations that resemble the original. A positive weight moves the surrogate toward the explained class and a negative weight moves it away. The weights describe the surrogate near one input only; they are neither global importance measures nor proof that the model verified a claim.

5.3 Inference and Review Workflow

The operational output is a review record rather than a bare label. For each submitted article, the classifier produces both softmax scores and selects the larger as its prediction. When an explanation is requested, the original text remains fixed as the reference instance while LIME queries perturbed versions through the same tokenizer and classifier. LIME does not change the trained parameters and is not an additional validation stage. Because repeated explanations may vary with the perturbation sample, an auditable implementation should record the model checkpoint, preprocessing version, predicted scores, explanation settings, and the text presented to the reviewer.

The prototype interface is implemented with Streamlit. A text-entry area accepts a news article and the interface returns the fake or real prediction, both softmax scores, and a token-level explanation, keeping the complete decision record visible in one place. The interface shows both class scores rather than only the winning label, retains the original text beside the highlighted tokens, and states that the scores are uncalibrated and advisory. It is a presentation layer for research and screening, not an unsupervised moderation system.

6. EXPERIMENTAL SETUP

6.1 Training Configuration

A pretrained BERT encoder and linear classification head are fine-tuned with categorical cross-entropy. AdamW performs parameter updates [16], a warm-up phase is followed by learning-rate decay, checkpoints are saved, validation performance is monitored, and mixed-precision computation is used [17]. The implementation uses Python, PyTorch, and Hugging Face Transformers in a CUDA-capable GPU/CPU environment. Table 4 summarizes the configuration. Numerical hyperparameters (BERT checkpoint, learning rate, batch size, number of epochs, maximum sequence length, warm-up ratio, weight decay, and random seed) and the split ratios are not reported in this version, which limits exact reproduction.

Table 4. Experimental Configuration

Component	Setting
Task	Binary article classification
Encoder	Pretrained BERT; all layers fine-tuned
Tokenizer	WordPiece with padding and truncation masks
Classifier	[CLS] linear head and softmax
Loss	Categorical cross-entropy
Optimizer	AdamW; warm-up then learning-rate decay
Computation	Mixed precision; CUDA-capable GPU/CPU
Model selection	Validation monitoring with checkpointing
Explanation	LIME local token attribution

AdamW decouples weight decay from the adaptive gradient update, while warm-up limits abrupt early changes to pretrained parameters. Mixed precision reduces memory use by performing selected operations at lower precision while retaining full precision where needed for stability. The test subset is held out from training and validation.

6.2 Evaluation Metrics and Statistical Analysis

Threshold-based performance is reported with accuracy, class-wise precision, recall (equal to sensitivity for the fake class and specificity for the real class), F1, balanced accuracy, Cohen's kappa, and the Matthews correlation coefficient (MCC). MCC is included because it uses all four cells of the confusion matrix and remains informative when class sizes differ [18]. Ranking performance is reported with ROC-AUC.

Point estimates on a finite test set carry sampling uncertainty, which becomes critical when the observed error count is zero. Exact Clopper-Pearson binomial intervals [19] are therefore computed for accuracy and for each class-wise rate. With zero observed errors in n items, the two-sided 95% interval for the error rate is $[0, 1 - 0.025^{1/n}]$, which is close to the well-known "rule of three" bound 3/n [20]. In addition, the probability of observing zero errors is computed as a function of the true error rate e, namely $(1 - e)^n$, to show which error rates the observed result can and cannot rule out. Finally, an exact two-sided binomial test checks whether the class mix of the test subset is consistent with that of the full corpus.

6.3 Evaluation Scenarios

Different split designs answer different questions. Table 5 lists the evaluation scenarios considered in this study, what each one measures, and its status. Only the in-domain random split is evaluated experimentally. The statistical analyses in Sections 7.2 and 7.3 re-examine that single test result and involve no new experiment. The cross-dataset, source-shortcut, temporal, and source-disjoint scenarios are not evaluated in this study: Table 2 provides literature context only, and these scenarios are specified as the protocol for follow-up work.

7. RESULTS AND DISCUSSION

7.1 Held-Out Test Performance

The held-out confusion matrix (Figure 3a) contains 245 fake and 255 real articles with zero off-diagonal entries, so all 500 test articles are classified correctly. The reported ROC-AUC is 1.00 (Figure 3b), which means that every real article received a higher real-class score than every fake article. The per-article prediction scores were not available for this revision, so the ROC curve itself is not reproduced; Figure 3b reports the AUC value only. Table 6 reports the complete set of metrics together with exact 95% confidence intervals.

Table 5. Evaluation Scenarios and Their Status in This Study

Scenario	Question answered	Status in this study
In-domain random split	How well does the model fit the corpus?	Directly evaluated (Section 7.1)
Statistical reliability	Which error rates does the result rule out?	Analysis of the Section 7.1 result; no new experiment

Scenario	Question answered	Status in this study
Test-subset class balance	Do test and corpus class proportions differ?	Class-proportion check only (Section 7.3)
Cross-dataset	Does performance transfer to other corpora?	Not evaluated; literature context only (Table 2)
Source shortcut / dateline ablation	Does the model rely on publisher cues?	Not evaluated; proposed protocol
Temporal and source-disjoint	Does it generalize to new events and outlets?	Not evaluated; proposed protocol (Section 10)

Because there are no false positives or false negatives, every threshold-based metric equals 1.00, including the chance-corrected measures (kappa and MCC). This agreement is internally consistent: zero off-diagonal counts imply perfect thresholded metrics, and complete separation of scores implies an ROC-AUC of 1.00. The support counts are close, so accuracy is not dominated by one class, and the balanced accuracy equals the plain accuracy.

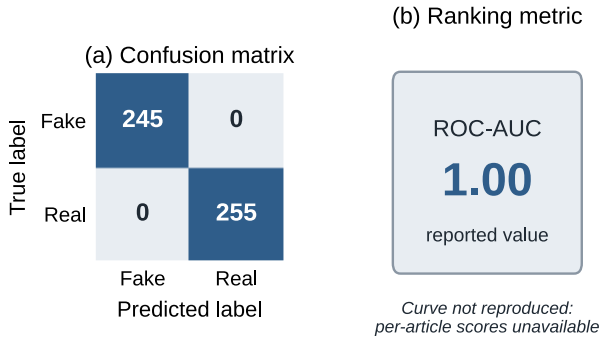


Figure 3. Held-out test results: (a) confusion matrix and (b) reported ROC-AUC (the ROC curve is not reproduced because per-article scores were unavailable).

Table 6. Held-Out Test Metrics with Exact 95% Clopper-Pearson Confidence Intervals

Measure	Fake	Real	Overall (95% CI)
Support (true class size)	245	255	500
Precision	1.00	1.00	–
95% CI for precision (n = predicted count)	0.985–1	0.986–1	–
Recall	1.00	1.00	–
95% CI for recall (n = class support)	0.985–1	0.986–1	–
F1	1.00	1.00	1.00
Accuracy	–	–	1.00 (0.993–1)
Balanced accuracy	–	–	1.00
Cohen's kappa	–	–	1.00
MCC	–	–	1.00
ROC-AUC (reported)	–	–	1.00
In-domain error-rate upper bound	–	–	0.6% (one-sided 95%, binomial assumption)

7.2 Statistical Reliability of the Perfect Score

A perfect score on 500 items does not establish a zero error rate. Figure 4 shows the exact 95% confidence intervals: overall accuracy lies between 0.993 and 1.000, and each class-wise recall or precision lies between about 0.985 and 1.000. For recall, the interval uses the true class size as denominator (245 fake, 255 real); for precision, it uses the number of articles predicted as that class, which here equals the class size because there are no errors. The class-wise intervals are wider than the overall interval because each is based on roughly half of the test set.

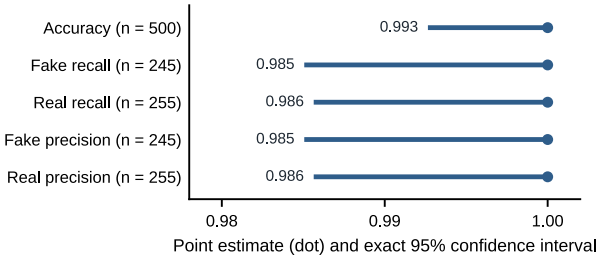


Figure 4. Exact 95% Clopper-Pearson confidence intervals for the test metrics.

Figure 5 and Table 7 express the same uncertainty in terms of the true error rate. If the classifier's true error rate were 1%, the probability of seeing no errors in 500 articles would be only 0.7%, so such an error rate is effectively ruled out. If the true error rate were 0.5%, the probability would be 8.2%, and for 0.1% it would be 60.6%; these error rates remain fully consistent with the observation. The one-sided 95% upper bound on the error rate is therefore $1 - 0.05^{1/500} \approx 0.597\%$, or about 0.6%. This bound assumes that the 500 test articles are independent draws from the same distribution as the evaluated corpus, so that the error count follows a binomial distribution. It describes the in-domain error rate only. It does not imply a 0.6% error rate for other domains, publishers, time periods, or deployment conditions, which were not evaluated. Narrowing the in-domain bound to 0.1% would require about 3,000 error-free test articles, so a larger held-out set is recommended for future reporting.

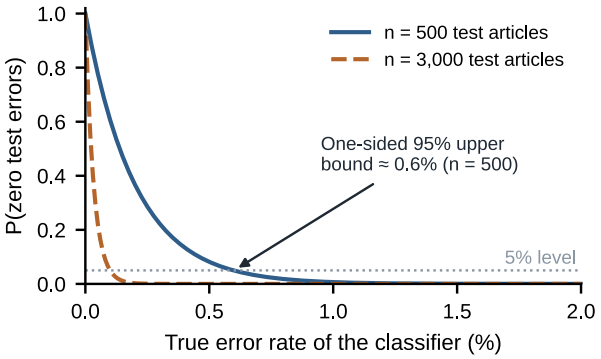


Figure 5. Probability of observing zero test errors as a function of the true error rate, for the current test size (500) and a larger test set (3,000).

Table 7. Zero-Error Sensitivity Analysis (n = 500)

True error rate	P(0 errors in 500 articles)
0.1%	0.606
0.2%	0.368
0.5%	0.082
0.6%	0.049
1.0%	0.007
2.0%	< 0.001

7.3 Class Proportions of the Test Subset

The fake share of the test subset is 49.0% (245 of 500), compared with 52.3% in the full corpus. An exact two-sided binomial test gives p = 0.14, so no statistically significant difference in class

proportions was detected. This check concerns class proportions only; it does not show that the test subset is representative of the corpus in topics, sources, or publication time, which were not examined. A stratified split would remove the class-proportion variation and is recommended for future runs.

7.4 Why Separation May Be Easy on This Corpus

The perfect separation is compatible with genuine semantic discrimination, but three documented properties of ISOT, with which the corpus composition is consistent, offer simpler explanations. First, all ISOT real articles come from one wire service and share a distinctive house style, including the "(Reuters)" dateline near the start of the text, which falls inside the truncated encoder window. Second, fake articles come from a limited set of sites flagged as unreliable, which may carry their own recurring vocabulary, formatting, and attribution patterns. Third, if the articles cover a narrow publication period, topics and named entities are shared between training and test articles. Exploratory global token-attribution plots inspected during model development were observed to give prominence to terms such as Reuters, Kuala Lumpur, and regional entity names. Those plots are not reproduced in this paper and the observation was not quantified, so it is offered only as a hypothesis that source and topic signatures contribute to the decision, to be tested by the dateline-ablation scenario in Table 5.

These properties do not invalidate the result, but they bound its interpretation: the model may have learned to separate the corpus's two source populations, which is a necessary but not sufficient condition for detecting misinformation in general. The cross-dataset contrast in Table 2 is consistent with this reading, and the source-disjoint, temporal, cross-dataset, and dateline-ablation scenarios in Table 5 are designed to measure how much of the performance survives when these signatures are removed.

7.5 Probability Interpretation

Softmax values rank the two classes and support threshold analysis, but high confidence alone does not demonstrate calibration; modern neural networks are often over-confident [14]. The confidence scores produced on the test set are concentrated near one, which is expected for a perfectly separated sample but says little about confidence on unfamiliar sources. Before operational use, scores should be calibrated on independent validation data, for example with temperature scaling, and low-confidence or distribution-shifted articles should be routed to human review.

8. EXPLAINABILITY ANALYSIS

8.1 Local Attribution Result

Figure 6 shows a LIME explanation for the short input "take unplan pto order travel home town." The classifier assigns 0.82 to class 0 (fake) and 0.18 to class 1 (real). The token "unplan" contributes most strongly toward the fake class, while "pto," "take," "order," "travel," "home," and "town" contribute toward the real class with smaller weights.

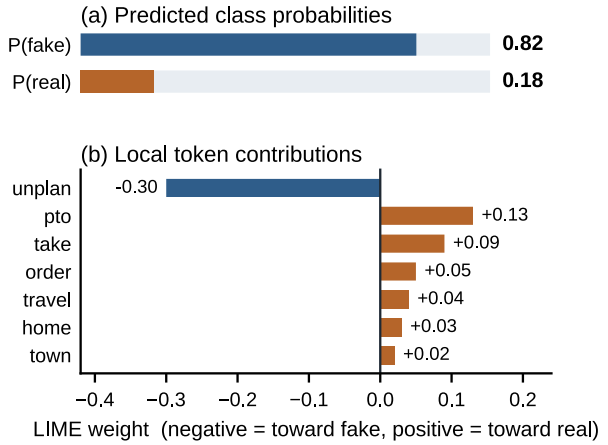


Figure 6. LIME output for one input: (a) predicted class probabilities and (b) local token contributions.

Table 8 quantifies how the attribution is distributed. A single token, "unplan," accounts for 45.5% of the total absolute attribution, and the top two tokens account for 65.2%. The six real-leaning weights sum to +0.36, slightly more than the single fake-leaning weight of −0.30, so the net token contribution (+0.06) points toward the real class even though the prediction is fake. The surrogate therefore attributes most of the fake prediction to its intercept, that is, to the baseline behavior of the model in this neighborhood rather than to the displayed tokens. This is a typical property of short, low-content inputs and shows why a single explanation should not be read as a complete account of a decision.

Table 8. Distribution of LIME Attribution for the Example in Figure 6

Token	Weight	Direction	Share of total
unplan	−0.30	Fake	45.5%
pto	+0.13	Real	19.7%
take	+0.09	Real	13.6%
order	+0.05	Real	7.6%
travel	+0.04	Real	6.1%
home	+0.03	Real	4.5%
town	+0.02	Real	3.0%
Net sum (signed)	+0.06	Real	—

8.2 Interpretation Boundaries

A locally faithful explanation can still describe an undesirable rule. If removing a publisher name changes the prediction, LIME exposes that sensitivity without determining whether reliance on the name is justified. The exploratory global plots mentioned in Section 7.4, in which proper nouns and fragmentary tokens were prominent, are more consistent with source or topic signatures than with universal indicators of falsehood; because they are not reproduced or quantified here, this remains a qualitative observation. It nevertheless argues against assuming that the model has learned general semantic truth cues, and it illustrates the value of explanation as an audit tool that can reveal shortcuts a perfect test score would conceal.

8.3 Explanation Validation Protocol

Explanation quality requires an evaluation separate from predictive accuracy. A fidelity check removes or masks the most influential tokens and records whether the target probability changes in the direction suggested by the surrogate. A stability check repeats the explanation with different random seeds and compares the highest-weight tokens. A contrastive check

compares similar articles assigned different labels. The audit sample should cover both predicted classes, high- and low-confidence items, different document lengths, and sources or time periods absent from training, and reviewers should label whether explanations emphasize claim content, stylistic artifacts, named entities, or publisher cues. Combining such attribution review with external test sets makes the explanation module a model-audit instrument rather than a decorative interface element.

9. LIMITATIONS AND RESPONSIBLE USE

Several limitations follow from the evidence. First, the model uses text only and ignores images, video, publication metadata, propagation networks, and external fact evidence. Second, the corpus, whose composition is consistent with ISOT, likely has a single-source real class and possibly a narrow time window, so it may encode source and topic regularities and does not reflect evolving misinformation. Third, results are reported for a single random split, numerical hyperparameters are not reported, and no variation across seeds is reported; repeated seeds, stratified splits, and larger test sets are needed to quantify run-to-run variation. Fourth, the softmax confidence is not calibrated. Fifth, LIME is local and perturbation-dependent, and token weights do not establish causality. Finally, because the test matrix contains no errors, a conventional analysis of false positives and false negatives cannot be performed; error analysis requires the stress scenarios in Table 5.

Automated labels can harm legitimate reporting if treated as final decisions. The system should therefore remain a decision-support tool with human oversight. Thresholds should reflect the relative costs of missing harmful content and of flagging legitimate reporting, and articles near the threshold, items from unseen domains, and cases with unstable explanations should receive manual review. Non-standard English or unfamiliar named entities must not become proxies for falsity, and reviewers need access to the original article and external evidence, not only the model explanation. Versioned models, documented data lineage, and periodic bias and drift audits are required if predictions affect moderation or publication decisions.

10. FUTURE WORK

The most important next step is to execute the evaluation scenarios in Table 5 with the same pipeline: a stratified test set of at least 3,000 articles with repeated seeds; deduplication before splitting; a dateline-ablation test that removes the "(Reuters)" prefix and location headers; temporal and source-disjoint splits; and cross-dataset transfer to FakeNewsNet and LIAR. A TF-IDF linear baseline and at least one alternative neural baseline should be trained on identical partitions so that the contribution of BERT can be quantified.

Further directions include calibration on held-out data, comparison of LIME with SHAP [11] and transformer-specific attribution methods under identical perturbation budgets, multimodal extensions that combine the article encoder with visual and metadata representations, and multilingual evaluation with language-appropriate encoders. Finally, a user study should test whether the combined label, scores, and token evidence improve reviewer triage accuracy or reduce unwarranted trust compared with a label-only interface.

11. CONCLUSION

This paper presented an explainable transformer-based fake news detector in which WordPiece inputs are encoded by BERT, the [CLS] representation is mapped to fake and real probabilities, and LIME exposes local token contributions. On a 500-article held-out subset of a merged fake/real news corpus whose composition is consistent with the ISOT Fake News Dataset, the classifier achieved perfect precision, recall, F1, accuracy, MCC, and ROC-AUC. Under a binomial sampling assumption, this result bounds the in-domain error rate below 0.6% with 95% confidence, a bound that does not extend to other domains or deployment conditions; no statistically significant difference between test-subset and corpus class proportions was detected. Literature comparison and qualitative explanation analysis suggest that the corpus is separable partly through source and topic signatures, so the result supports strong in-domain performance rather than general misinformation detection. The explanation module is valuable because it can surface such signatures. Within these limits, the BERT-LIME system provides a transparent baseline, and the evaluation protocol defined here sets out the evidence needed for broader claims.

12. REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] 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.

[3] 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, 2017, pp. 422–426.

[4] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," Big Data, vol. 8, no. 3, pp. 171–188, 2020.

[5] 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.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[7] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," Nature Machine Intelligence, vol. 2, pp. 665–673, 2020.

[8] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proceedings of NAACL-HLT, 2016, pp. 1480–1489.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008.

[10] 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, pp. 11765–11788, 2021.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural

Information Processing Systems, vol. 30, 2017, pp. 4765–4774.

[12] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC 2017), Lecture Notes in Computer Science, vol. 10618, Springer, 2017, pp. 127–138.

[13] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," Security and Privacy, vol. 1, no. 1, e9, 2018.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1321–1330.

[15] M. H. Goldani, S. Momtazi, and R. Safabakhsh, "Detecting fake news with capsule neural networks," arXiv preprint arXiv:2002.01030, 2020.

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proceedings of the International Conference on Learning Representations (ICLR), 2019.

[17] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[18] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, art. 6, 2020.

[19] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," Biometrika, vol. 26, pp. 404–413, 1934.

[20] J. A. Hanley and A. Lippman-Hand, "If nothing goes wrong, is everything all right? Interpreting zero numerators," JAMA, vol. 249, no. 13, pp. 1743–1745, 1983.