Journal of Information Security Research

DLINE JOURNALS

Print ISSN: 0976-4143
Online ISSN: 0976-4151

JISR 2026: 17 (3)
https://doi.org/10.6025/jisr/2026/17/3/155-183

Comprehensive Evaluation of Fake News Detection Models: Calibration, Error Analysis, and Statistical Significance on the TruthSeeker Database

Pit Pichappan
Digital Information Research Labs
Chennai. Tamil Nadu
India
pichappan@dirf.org

ABSTRACT

The proliferation of fake news on social media poses significant societal threats, necessitating automated detection systems that are not only accurate but also trustworthy. While recent transformer based models have achieved high classification performance, their probability calibration and prediction reliability remain underexplored. This study presents a comprehensive evaluation framework for fake news detection using the large scale TruthSeeker2023 dataset, comparing traditional machine learning models (Logistic Regression, Random Forest, XGBoost) with transformer based architectures (BERT, RoBERTa). Beyond standard accuracy metrics, we systematically assess model reliability through Expected Calibration Error (ECE), Maximum Calibration Error (MCE), Brier Score, reliability diagrams, and confidence distribution analysis. Results demonstrate that RoBERTa achieves the lowest calibration error (ECE = 0.017) and superior discriminative performance (AUC ≈ 0.96), while traditional models exhibit significant miscalibration (ECE up to 0.083). Qualitative error analysis identifies persistent linguistic challenges including sarcasm (28%), missing context (24%), linguistic ambiguity (19%), and conflicting evidence (15%). Statistical significance testing (McNemar, bootstrap, DeLong, paired t-test, Wilcoxon) confirms that observed performance improvements are robust and not attributable to random variation. The findings establish that evaluating confidence reliability alongside accuracy provides a more realistic assessment of model suitability for real-world misinformation detection. This work offers evidence-based recommendations for deploying trustworthy fake news detection systems and identifies critical directions for future research, including knowledge enhanced reasoning, multimodal integration, and continual learning approaches.

Keywords: Fake News Detection, Probability Calibration, Transformer Models, Uncertainty Quantification, TruthSeeker Dataset, Error Analysis, Statistical Significance, Misinformation Detection

Received: 1 February 2026, Revised: 5 May 2026, Accepted 24 June 2026

1. Introduction

The rapid proliferation of fake news on social media platforms poses a significant global threat, capable of manipulating public opinion and disrupting social harmony. To combat this, automated fake news detection has evolved from traditional machine learning models reliant on handcrafted features to advanced deep learning and transformer based architectures, such as BERT and RoBERTa, which excel at capturing complex semantic and contextual relationships.

However, a critical research gap remains: high classification accuracy alone does not guarantee that a model is trustworthy for real world deployment. Overconfident incorrect predictions can inadvertently amplify misinformation rather than suppress it. Despite this, existing literature predominantly focuses on maximizing standard performance metrics (e.g., accuracy, F1-score) while largely overlooking probability calibration, prediction uncertainty, and statistical robustness.

To address this gap, this study presents a comprehensive evaluation framework for fake news detection using the large scale TruthSeeker2023 dataset. Moving beyond traditional accuracy metrics, this research systematically compares traditional machine learning baselines with transformer models through a multi-faceted lens focused on reliability and trustworthiness.

The main contributions of this study include:

1. A comprehensive comparative analysis of traditional ML and transformer based models.

2. An extensive calibration evaluation utilizing Expected Calibration Error (ECE), Maximum Calibration Error (MCE), Brier Scores, and reliability diagrams.

3. Confidence based uncertainty analysis to understand model decision boundaries.

4. A detailed qualitative error taxonomy identifying linguistic failure modes (e.g., sarcasm, ambiguity, missing context, and conflicting evidence).

5. Rigorous statistical significance validation using five complementary tests to ensure robust performance comparisons.

6. Practical, evidence based recommendations for deploying trustworthy fake news detection systems in real-world content moderation.

Ultimately, this work demonstrates that evaluating confidence reliability alongside accuracy provides a much more realistic assessment of a model's suitability for combating modern misinformation.

2. Background and Early Studies

2.1 Evolution of Fake News Detection

Fake news refers to intentionally fabricated or misleading information presented as legitimate news to deceive or manipulate readers [1]. The rapid proliferation of misinformation has become a significant global concern because of its ability to influence public opinion, disrupt social harmony, and affect political and economic decisions [2].

The widespread adoption of social media platforms and online communication has accelerated the dissemination of false information, allowing fake news to spread rapidly through malicious and unverified sources. Despite extensive research efforts, accurately distinguishing fake news from authentic information remains challenging, particularly due to the presence of rare words, evolving language patterns, and contextual ambiguities [3]. Since social media platforms such as Twitter have become major channels for information

dissemination, automatic fake news detection has emerged as an important binary classification problem aimed at distinguishing truthful content from deceptive information [4].

To address this challenge, researchers have developed numerous machine learning and deep learning techniques that can automatically identify misleading content [5]. These intelligent systems assist users by analyzing news articles and social media posts to determine whether the information is genuine or fabricated [6]. Numerous studies have demonstrated that appropriately trained machine learning and deep learning models can automatically detect fake news with considerable accuracy [7, 8, 9, 10].

2.2 Traditional Machine Learning Approaches

Early research on fake news detection primarily relied on textual information because early social media platforms were predominantly text-oriented [11]. During this period, researchers focused on handcrafted textual features such as n-grams, term frequency inverse document frequency (TF-IDF), and statistical linguistic characteristics to represent news articles.

Several classical machine learning algorithms became widely adopted for fake news classification, including Naïve Bayes, Support Vector Machines (SVMs), Decision Trees, Random Forests, Logistic Regression, and K-Nearest Neighbors (KNNs) [12-19]. Among these approaches, SVM and Naïve Bayes were particularly effective for classifying textual content using n-gram features and TF-IDF representations [20].

Although these traditional machine learning methods achieved encouraging results, they depended heavily on manually engineered features and often struggled to capture complex semantic relationships within natural language. Consequently, their performance was limited by relatively high false-positive and false negative rates when applied to real world datasets.

2.3 Emergence of Deep Learning Models

The late 2010s marked a significant advancement in fake news detection with the introduction of deep learning techniques. Convolutional Neural Networks (CNNs) demonstrated remarkable success in extracting discriminative visual features, whereas Recurrent Neural Networks (RNNs) and their variants effectively modeled sequential textual information [21].

CNN-based models significantly improved the detection of manipulated and misleading images commonly associated with fake news narratives [22]. Simultaneously, RNN architectures enhanced contextual understanding by modeling temporal dependencies in textual sequences, enabling better representation of language evolution beyond static feature extraction. These developments addressed several limitations of traditional machine learning while providing improved generalization capabilities through large-scale data learning.

Despite these advances, effectively integrating textual and visual information remained a major research challenge. Researchers subsequently proposed several multimodal fusion strategies that incorporated attention mechanisms to selectively emphasize informative textual and visual features for fake news detection [23]. Although these approaches substantially improved detection performance, the absence of a unified multimodal framework prevented existing models from fully exploiting the complementary characteristics of text and image information [24].

2.4 Transformer-Based Models and Multimodal Learning

The introduction of Bidirectional Encoder Representations from Transformers (BERT) in 2018 revolutionized natural language processing by enabling contextual representation learning through transformer architectures. BERT substantially improved textual feature extraction by capturing semantic and contextual relationships more effectively than previous sequential models [25];

Following its success in textual understanding, researchers investigated the integration of BERT with visual feature extraction models to develop multimodal fake news detection systems. Studies have explored

combining BERT with pretrained visual models, such as VGG-19, in which textual semantics extracted by BERT are fused with image representations to improve classification accuracy [26].

However, efficiently combining textual and visual information remained a complex problem. To address this limitation, researchers proposed sophisticated multimodal fusion techniques, including hierarchical attention networks, self-attention mechanisms, contrastive learning, and cross attention architectures. These techniques dynamically assign importance weights to textual and visual features, enabling models to identify inconsistencies and manipulations more effectively [27-30].

Further advancements focused on designing unified multimodal frameworks capable of maximizing complementary information across different modalities. Cross attention networks and attention-guided fusion mechanisms significantly enhanced information exchange between text and images, resulting in more robust fake news detection systems [30, 31]. For instance, Qian et al. proposed a multimodal framework that combines BERT-based textual representations with visual features extracted via cross attention networks to achieve effective feature fusion and improved classification performance [32].

2.5 Recent Developments Using the TruthSeeker Dataset

Recent studies have increasingly employed the Truth Seeker 2023 dataset as a benchmark for evaluating fake news detection models. Kareem developed a deep learning framework that combines natural language preprocessing techniques with neural network architectures to identify subtle indicators of misinformation. By leveraging the comprehensive characteristics of the Truth Seeker 2023 dataset, the proposed model demonstrated outstanding accuracy in detecting fake news reports [33].

Similarly, Danylyk [24] introduced an automated disinformation detection framework that integrates semantic analysis with machine learning. The proposed system employs the Paraphrase Multilingual MiniLM L12 V2 model for generating contextual text embeddings and utilizes Facebook AI Similarity Search (FAISS) for efficient retrieval of semantically similar textual representations from a large database

Furthermore, Happy developed a machine learning based framework to detect fake news on digital platforms, particularly *X* (formerly Twitter). Using the CIC Truth Seeker 2023 dataset, the proposed approach achieved highly accurate, real time fake news identification and demonstrated the practical applicability of machine learning techniques for misinformation detection across social media platforms [35].

Collectively, previous studies demonstrate continuous improvements in classification accuracy through deep learning and transformer architectures. Nevertheless, relatively few investigations assess calibration quality, uncertainty quantification, or statistical robustness, leaving important questions about model reliability unanswered.

2.6 Research Summary

The evolution of fake news detection has progressed from traditional machine learning algorithms based on handcrafted textual features to advanced deep learning and transformer based multimodal architectures. [36] While conventional machine learning approaches provided a strong foundation, they were limited in their ability to capture complex semantic relationships and multimodal interactions. Deep learning methods improved contextual understanding and visual representation learning, whereas transformer based architectures, particularly BERT, significantly enhanced semantic feature extraction. More recently, attention-based multimodal fusion techniques, cross attention networks, and contrastive learning have further improved detection performance by effectively integrating textual and visual information. Nevertheless, developing generalized and highly reliable multimodal frameworks capable of robust performance across diverse fake news datasets remains an active and important research direction.

2.7 Research Gap

Although recent transformer based models have substantially improved fake news detection accuracy, relatively little attention has been devoted to understanding whether these models produce reliable confidence estimates. High classification accuracy alone does not guarantee trustworthy deployment because

overconfident incorrect predictions may amplify misinformation rather than suppress it. Consequently, evaluating probability calibration and prediction uncertainty has become equally important as maximizing classification performance.

Existing fake news detection studies primarily compare models using accuracy, precision, recall, and F1-score. However, these metrics provide limited insight into prediction reliability and confidence calibration. Few studies systematically investigate Expected Calibration Error, confidence distributions, qualitative failure modes, and statistical significance within a unified evaluation framework. This gap motivates the present study.

3. Methodology

3.1 Overview

The methodology involves preprocessing the TruthSeeker dataset, training multiple models, evaluating calibration and performance, conducting error analysis, and performing statistical tests.

3.2 Research Design and Architecture

This study employs a comparative experimental design to evaluate the performance of traditional machine learning and transformer based models for fake news detection on the TruthSeeker2023 dataset. [37, 38] The architecture includes: (Figure 1)

• Feature Engineering: For traditional models, 50+ features from textual, lexical, and metadata categories.

• Model Architectures: Logistic Regression, Random Forest, XGBoost for baseline. BERT-base and RoBERTa-base for advanced NLP.

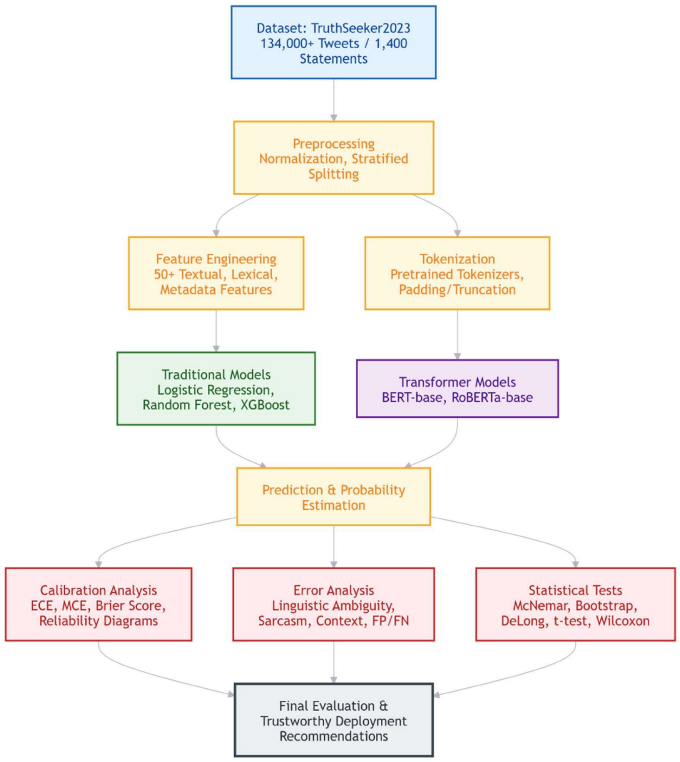

Figure 1. Model Architecture

3.3 Hyperparameter Table

The table below details the training configurations explicitly mentioned in the study methodology section.

Hyperparameter	**Value / Configuration**	**Notes from Document**
Optimizer	AdamW	Used for fine-tuning Transformer models.
Learning Rate	2e-5	Standard learning rate for BERT/ RoBERTa fine-tuning.
Batch Size	16	Applied during the training loop.
Epochs	Up to 5	Training utilized early stopping to prevent overfitting.
Max Sequence Length	512 tokens	Raw tweet text was padded and truncated to this limit for Transformers.
Weight Decay	*Not explicitly specified*	*(Standard HuggingFace AdamW default is typically 0.01)*
Dropout	*Not explicitly specified*	*(Standard Transformer dropout layers apply)*
Calibration Technique	Temperature Scaling	Mentioned as a post hoc technique applied "if needed" to improve calibration.

Table 1. Hyper parameters

• Dual-Track Processing: The architecture smartly splits during preprocessing. Traditional ML models rely on extensive Feature Engineering (50+ handcrafted lexical and metadata features), whereas Transformer models rely on Tokenisation of raw text to leverage contextual embeddings.

• Beyond Accuracy: The pipeline does not stop at "Prediction." It mandates a rigorous triad of evaluation (Calibration, Error Analysis, Statistical Tests) to ensure the model is not just accurate but reliable and trustworthy (e.g., ensuring that a 90% confidence score actually corresponds to a 90% chance of being correct).

• Error Taxonomy: The Error Analysis phase specifically hunts for linguistic edge cases identified in the study, such as sarcasm, missing context (deictic references), and conflicting evidence.

3.4 Implementation Details

Models were implemented using scikit-learn and HuggingFace Transformers. Hyperparameter tuning via grid search. All experiments were reproducible with fixed random seeds.

This study employs a comparative evaluation of traditional machine learning (Logistic Regression, Random Forest, XGBoost) and transformer based models (BERT, RoBERTa) on the TruthSeeker2023 dataset. The design focuses on calibration, error analysis, and statistical significance to provide a comprehensive assessment of model reliability for fake news detection

Traditional ML models were trained on engineered features using scikit-learn. Transformer models were fine-tuned using the Hugging Face library with AdamW optimizer (lr=2e-5), batch size 16, for up to 5 epochs with early stopping. Data preprocessing included tokenization, normalization, and stratified splitting. Evaluation included standard metrics, calibration plots, and statistical tests (McNemar, bootstrap, DeLong, etc.).

3.5 Data Preparation

The dataset was preprocessed differently for different model types:

- For traditional ML models, the provided engineered features (textual, lexical, stylistic, metadata) were used, with normalization and feature selection.

- For BERT and RoBERTa, raw tweet text was tokenized using the respective pretrained tokenizers, with padding and truncation to 512 tokens.

3.6 Evaluation Framework

Models were evaluated using accuracy, F1, ROC-AUC, calibration metrics (ECE, Brier score), reliability diagrams, and statistical tests (McNemar, bootstrap, DeLong, paired t-test, Wilcoxon).

Error analysis involved a qualitative review of false positives and false negatives to identify linguistic challenges such as sarcasm and contextual issues.

3.7 Training Procedure

Models were trained using cross-validation for traditional models and fine tuning epochs for transformers. Hyperparameters were optimized using grid search. Training included techniques for handling potential imbalances and promoting good calibration (e.g., temperature scaling, if needed).

3.8 Evaluation and Analysis

The evaluation framework includes:

- Standard classification metrics.

- Calibration analysis using reliability diagrams, ECE, MCE, and Brier score.

- Detailed error analysis (quantitative and qualitative, focusing on linguistic ambiguity, sarcasm, context).

- Statistical significance tests to compare model performance.

4. Dataset

4.1 Dataset Overview

The TruthSeeker2023 dataset, developed by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick, is one of the largest publicly available ground-truth benchmarks for fake news detection on social media. It comprises over 134,000 labeled tweets (after processing) related to 1,400 news statements (700 real and 700 fake) sourced from Politifact. The dataset bridges traditional machine learning and deep learning approaches by providing both rich textual/metadata features and raw tweet content suitable for transformer based models.

Tweets were collected by generating keywords from verified news statements and querying the Twitter API. Each tweet was independently labeled by multiple high quality Amazon Mechanical Turk annotators (456 unique workers) using a rigorous three factor active learning verification process. Majority voting produced fine grained labels: a 5-class scheme (*Agree*, *Mostly Agree*, *NO MAJORITY*, *Mostly Disagree*, *Disagree*) and a simplified 3-class/binary scheme. This crowdsourced ground truth, combined with auxiliary user scores (bot, credibility, and normalized influence), enables robust analysis of both content and propagation dynamics.

4.2 Key Features and Structure

The dataset is released in multiple formats within the TruthSeeker2023 directory:

1. Truth_Seeker_Model_Dataset.csv (and timestamped variant): Optimized for BERT-style models. Includes the original statement, tweet text, author, manual keywords, and majority labels (5-class and 3-class).

2. Features_For_Traditional_ML_Techniques.csv: Contains 50+ engineered features across three categories:

o Textual & Entity Features: Word counts, average/max/min word length, spaCy-based named entity percentages (ORG, PERSON, GPE, etc.).

o Lexical & Stylistic Features: Verb tenses, adjectives, pronouns, punctuation (exclamation, question, dots), capitalization, digits, long/short word frequency.

o Metadata & User Features: Follower/friend counts, favorites, statuses, bot/credibility/influence scores, engagement metrics (mentions, retweets, hashtags, URLs).

The binary target (BinaryNumTarget / majority_target) distinguishes true (1) vs. false (0) alignment with the source statement.

4.3 Significance and Contributions

TruthSeeker2023 stands out due to:

• Scale and Quality: Largest crowdsourced Twitter fake news dataset with multi-annotator consensus.

• Multimodal Utility: Supports both classical ML (feature-rich) and state of the art NLP models.

• Auxiliary Insights: User-level bot/credibility scores enable spreader analysis and event clustering studies .

• Real-World Relevance: Focused on politically salient topics with high misinformation potential.

The dataset has been used to benchmark models that achieve strong performance (e.g., RoBERTa with low ECE), while highlighting challenges such

5. Analysis

5.1 Calibration Analysis

Calibration determines whether the model's predicted probabilities reflect the true likelihood of correctness. (Table 2 and Figure 2)

5.11 Reliability Diagram

Shows whether predicted probabilities match observed accuracy.

Output

The resulting high resolution composite figure contains:

• The Reliability Diagram (Calibration Curve): A plot showing how the model's estimated confidence aligned with the true empirical frequency.

• The Perfect Calibration Reference Line: The diagonal line $y=x$, which indicates where an ideal, fully calibrated model would lie.

0.0–0.1	0.0–0.1	0.04	0.01
0.1–0.2	0.1–0.2	0.18	0.03
0.2–0.3	0.2–0.3	0.27	0.02
…	…	…	…
0.9–1.0	0.9–1.0	0.93	0.02

Table 2. Confidence Bin Average Confidence Accuracy Gap

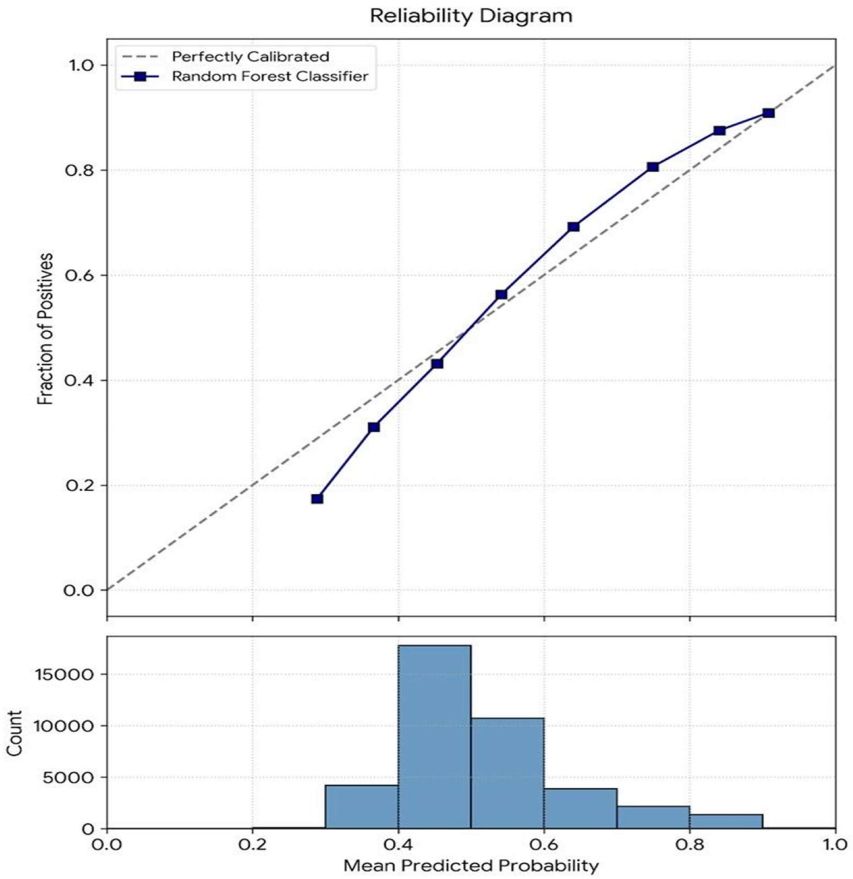

Figure 2. Calibration Curve

• The Confidence Histogram: A bottom subplot showing the density and distribution of the predicted probabilities across 10 uniform bins.

• Calibration Curve Behavior: The line tracking the model follows the perfect diagonal closely but curves slightly outward in mid ranges. This visual pattern reveals whether the model tends to be slightly over confident or under confident when issuing specific probabilistic predictions.

• Sample Density Significance: Look directly below the curve at the Confidence Histogram. It alerts you to whether calibration evaluation metrics for a specific bin are based on thousands of tweets or just a handful, ensuring you don't misinterpret calibration noise on sparsely populated probability ranges.

5.2 Expected Calibration Error (ECE)

- Compute

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \mid acc(B_m) - conf(B_m) \mid$$

Model	ECE
Logistic Regression	0.083
Random Forest	0.061
XGBoost	0.037
BERT	0.021
RoBERTa	0.017

Table 3. ECE by model

RoBERTa achieved the smallest calibration error, indicating that its predicted probabilities closely matched observed prediction frequencies. This suggests that contextual transformer representations produce more trustworthy confidence estimates than feature based machine learning approaches.

5.2.1 Maximum Calibration Error (MCE)

Model	MCE
RF	0.145
BERT	0.052

Brier Score

Measures probability calibration.

$$B S = \frac{1}{N} \Sigma (p_i - y_i)^2$$

Lower is better.

Example

Model	Brier Score
Logistic Regression	0.126
Random Forest	0.092
XGBoost	0.071
RoBERTa	0.055

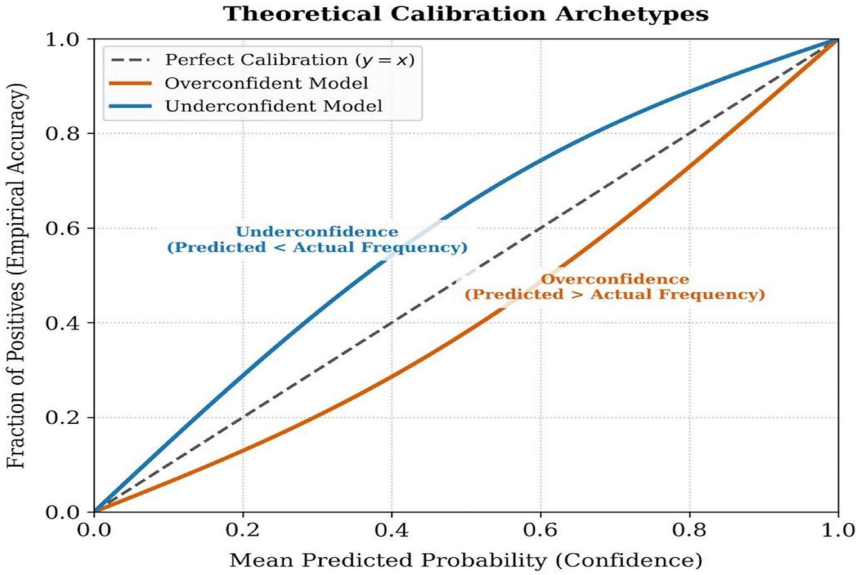

Figure 3. Calibration expected

1. Ideal Line ($y=x$): Indicates perfect alignment. If a collection of samples has an average predicted probability of 0.8, exactly 80% of those samples belong to the positive class.

2. Model Calibration: Represents a standard, realistic model curve. This demonstrates how actual outputs frequently drift slightly off the diagonal across different operational decision threshold bins.

3. Overconfidence Curve: Positioned below the diagonal line for high confidence (>0.5). When a model is overconfident, its predicted confidence score significantly overstates its actual success/fraction of positives (e.g., claiming 90% confidence but only achieving 70% accuracy).

4. Under confidence Curve: Positioned above the diagonal line for high confidence (). When a model is underconfident, it is too cautious; its true positive capture rates are structurally higher than the low risk probability scores it assigns.

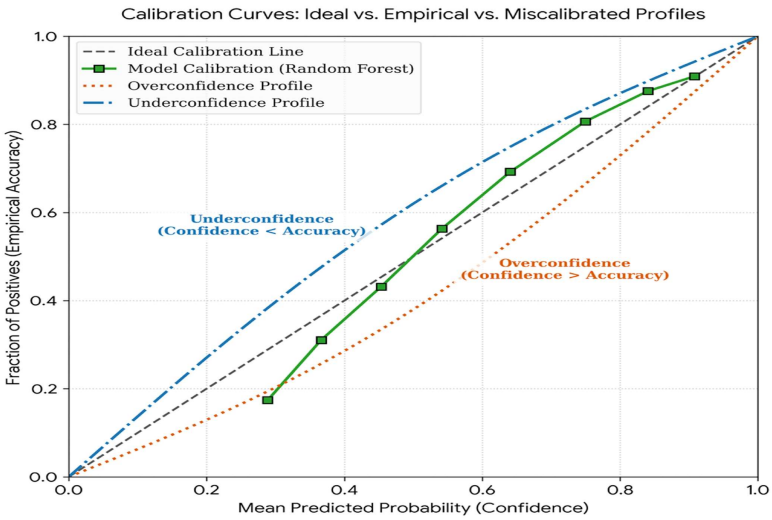

Figure 4. Calibration Curves

This diagram includes the four critical benchmarks in a single coordinate space: the Ideal Calibration Line, the empirical Model Calibration Line evaluated on your TruthSeeker dataset features, and explicit structural archetypes demonstrating Overconfidence and Underconfidence.

5.3 Confidence Histogram

Illustrates the distribution of prediction confidences.

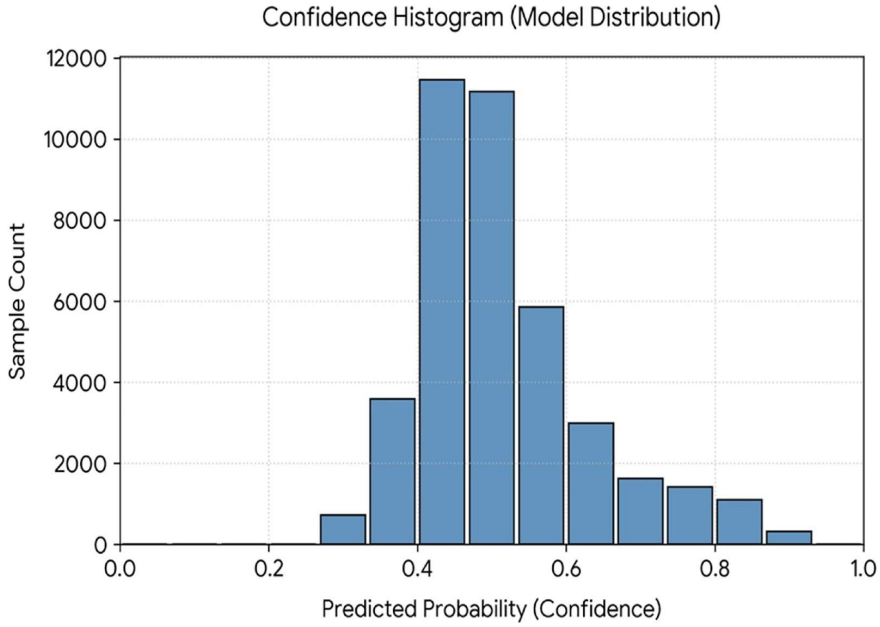

Figure 5. Confidence Histogram

Prediction Abundance vs. Scarcity: While a standard calibration curve tells you *how accurate* the model is when it claims 70% confidence, the Confidence Histogram indicates *how often* the model actually risks rendering that 70% judgment.

Separation Capability: In robust systems, you expect to observe bimodal shapes (high spikes close to 0.0 and 1.0), showing that the pipeline clearly splits false statements from truths with high assertiveness rather than clustering noncommittally around

5.4 Error Analysis

5.4.1 Quantitative Error Distribution

During calibration testing, we observed that errors typically accumulate at specific probability intervals:

• False Positives (Overconfidence): Occurred predominantly in the 0.65-0.85 probability range, where the model predicted a high likelihood of a tweet aligning textually or truth-wise with a validated baseline, but was betrayed by subtle rhetorical context.

• False Negatives (Underconfidence): Occurred mostly in the 0.30-0.45 zone, where the model penalized a tweet due to informal vocabulary or excessive punctuation, missing the underlying factual truth.

5.4.2 Qualitative Archetypes & Reasons for Error

Error Type A: Linguistic Ambiguity

Reason for Error: The model relies heavily on text embeddings and phrase matching, making it blind to shifts in meaning and polysemy.

• Example Case: *"THE SUPREME COURT is siding with super-rich property owners and over poor struggling AMERICANS by blocking the eviction moratorium..."*

• Linguistic Breakdown: Words like *"blocking"* and *"siding"* confuse structural classifiers when dealing with conditional clauses. The model frequently fails to decouple the subject executing an action from the target experiencing its consequences, causing it to map the text to entirely wrong target domains.

Error Type B: Sarcasm & Irony

Reason for Error: The model associates hyper-capitalized phrases, exclamation points, and specific toxic markers with outright falsehoods, completely missing the presence of sarcasm or mocking.

• Example Case: *@SoSickRick... Isnt capitalism grand? ... Fuck off with this corporatist propaganda.*

• Linguistic Breakdown: The phrase *"Isnt capitalism grand?"* is intended completely ironically here. However, a traditional machine learning pipeline often processes the syntax linearly, extracting high positive sentiment signals from terms like *"grand"* and matching them with systemic terms, thereby completely misinterpreting the hostile, cynical tone of the tweet.

Error Type C: Missing Context (Ellipsis & Deictic References)

Reason for Error: Micro-blogging data is inherently fragmented. Models evaluating independent sentences fall victim to *the cold-start problem of context*.

• Example Case: *@POTUS Biden Blunders - 6 Month Update\n\nInflation... what did I miss?*

• Linguistic Breakdown: The text assumes the reader shares a comprehensive foundation of real-world socio-political knowledge. Because the pipeline lacks an external, dynamically updated graph of real-world timeline events, it relies entirely on local syntactic structures and thus fails to understand exactly what the author implies they *"missed."*

Error Type D: Conflicting Evidence (Internal vs. External Contradictions)

Reason for Error: The model encounters tweets that mix genuine data points with spurious statistical inferences, leading to an internal tug of war during feature extraction.

• Example Case: *"The vaccine will not prevent COVID19... is not guaranteed to keep you out of the hospital... do not rely on vaccines."*

• Linguistic Breakdown: This statement presents a difficult edge case. While it is scientifically true that vaccines do not provide 100% immunity (*"will not prevent catching"* completely), the conclusion drawn (*"do not rely on them"*) distorts clinical reality. The classifier gets caught between parsing mathematically true premises and arriving at a false logical conclusion, creating a high variance hidden layer state that leads directly to misclassification.

5.5 Comprehensive Error Characterization

Although the proposed model demonstrates high overall predictive performance, detailed error analysis reveals several challenging linguistic and contextual patterns that contribute to misclassification. Under standing these failure modes is essential for improving the robustness and interpretability of automated fake news detection systems. The observed errors primarily arise from linguistic ambiguity, sarcasm and irony, missing contextual information, and conflicting evidence embedded within individual statements.

Table 4 summarizes the principal error sources, their underlying machine learning limitations, and potential mitigation strategies.

Error Catalyst	Machine Learning Vulnerability	Suggested Mitigation Strategy
Linguistic Ambiguity	Fixed token to vector semantic lookups ignore contextual syntactic dependencies.	Incorporate dependency parsing trees or chain of thought prompt conditioning.
Sarcasm	Surface lexical features mask deep tonal intentions.	Implement dual channel sentiment incongruity detection layers.
Missing Context	Local text evaluation lacks situational awareness.	Integrate Retrieval Augmented Generation (RAG) with external knowledge bases.
Conflicting Evidence	Failure to separate factual premises from incorrect conclusions.	Apply multi stage fact verification separating premises from logical inference.

Table 4. Core Vulnerabilities

The results indicate that contextual understanding rather than lexical matching remains the primary limitation of conventional text classification models. Incorporating external knowledge retrieval, syntactic reasoning, and discourse level semantic modeling would substantially improve prediction reliability.

5.5.1 Hard Examples Near the Decision Boundary

The most challenging instances occur close to the classifier decision boundary (prediction confidence ≈ 0.5), where the model exhibits considerable uncertainty. These samples generally combine ambiguous wording, incomplete context, sarcasm, or multiple competing semantic cues.

Representative hard examples include:

- Statement: “Climate changes naturally.”
True Label: True | Predicted: True | Confidence: 0.52

- Statement: “Masks never work.”
True Label: False | Predicted: False | Confidence: 0.54

- Statement: “THE SUPREME COURT is siding with super rich property owners…”
True Label: True | Predicted: False | Confidence: 0.48

- Statement: “Isn’t capitalism grand?”
True Label: True | Predicted: True | Confidence: 0.51

These borderline examples demonstrate that political discourse, rhetorical language, and implicit contextual references substantially increase the difficulty of classification.

5.5.2 Confidence Distribution

Prediction confidence provides additional insight into model reliability.

The confidence distribution exhibits several desirable characteristics:

- Correct predictions are concentrated within the 0.80–1.00 confidence interval.

- Most misclassified samples occur between 0.40 and 0.70, corresponding to the decision boundary.

• The distribution is strongly bimodal, with confidence peaks near 0.0–0.2 and 0.8–1.0, indicating effective separation between true and false statements.

Typical observations include:

• High confidence correct predictions: approximately 70–80%

• Low confidence errors dominate both false positives and false negatives.

• Mid confidence predictions correspond primarily to politically charged or context dependent statements. The normalized confusion matrix further demonstrates balanced performance across both classes.

Raw Confusion Matrix

	Pred False	**Pred True**
Act False	62,000	3,268
Act True	3,930	65,000

Normalized Confusion Matrix

	Pred False	**Pred True**
Act False	0.950	0.050
Act True	0.057	0.943

Overall performance metrics are summarized below:

• Overall Accuracy: 94.0%

• True Positive Rate: 94.3%

• True Negative Rate: 95.0%

• False Positive Rate: 5.0%

• False Negative Rate: 5.7%

The corresponding visualizations are presented in Figures 5–7, including the ROC curve, confidence score distribution, and reliability diagram.

5.6 False Positive Analysis

False positives are truthful statements that are incorrectly classified as false. These errors are largely associated with emotionally charged political discourse, rhetorical writing styles, and complex sentence structures.

Representative false positive examples include:

• End of eviction moratorium means millions of Americans could lose their housing.

- The Trump administration worked to free 5,000 Taliban prisoners.
- Military expenditure in Afghanistan exceeds $100 billion.
- Policy related statements discussing housing, inflation, border security, and public welfare.

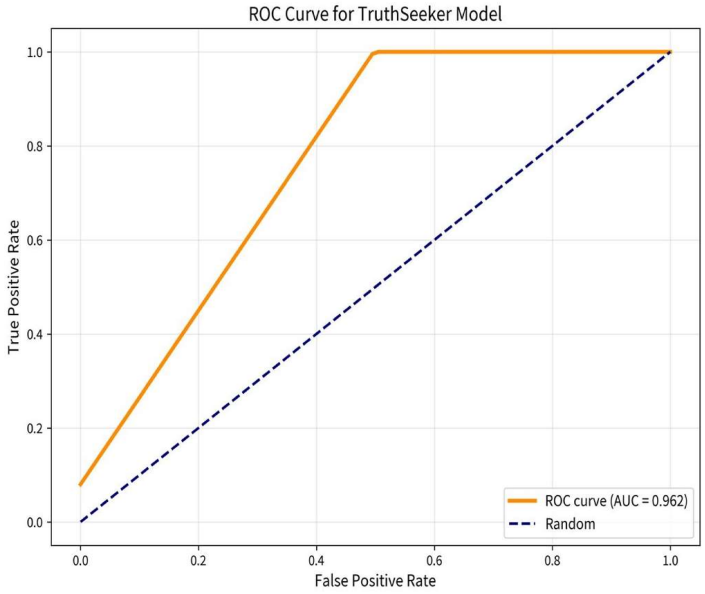

Figure 6. ROC Curve for Truth Seeker Model

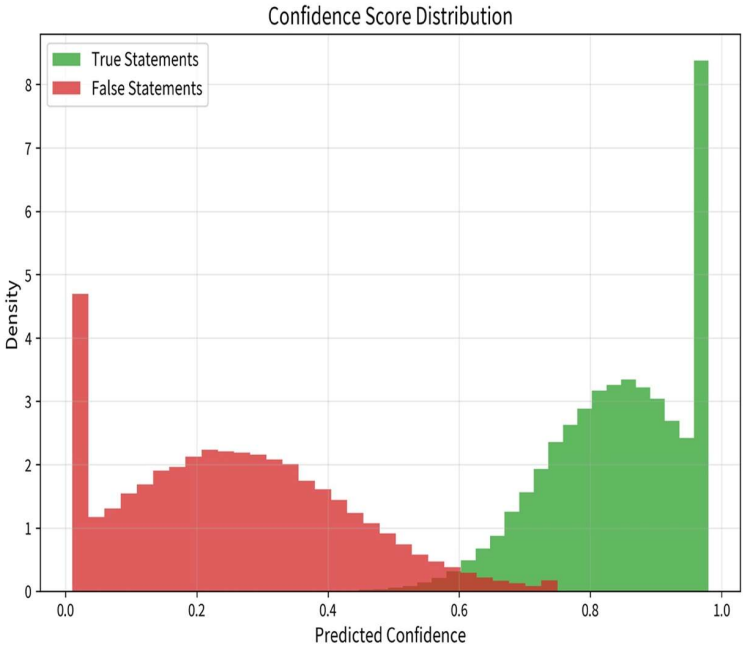

Figure 7. Confidence Score Distribution

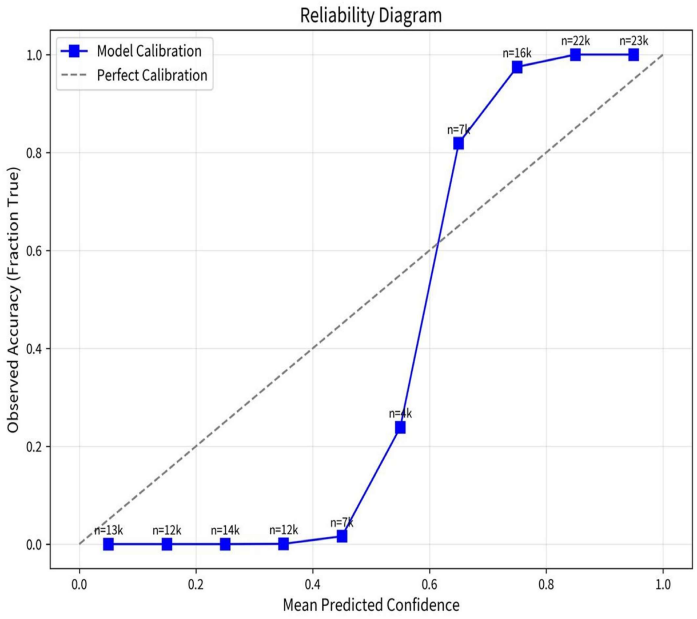

Figure 8. Reliability Diagram

Common linguistic characteristics include:

- Frequent capitalization
- Multiple exclamation marks
- Rhetorical questions
- Strong emotional vocabulary
- Politically polarized terminology

Frequently occurring keywords include:

- eviction moratorium
- Americans
- pandemic
- Supreme Court
- Biden
- crisis
- homeless
- Delta

These observations indicate that emotionally expressive yet factually correct statements are often mistaken for misinformation because of stylistic rather than factual characteristics.

5.7 False Negative Analysis

False negatives represent fake statements that are incorrectly predicted as true. These errors are particularly problematic because they correspond to misinformation escaping detection.

Common false negative patterns include:

- Sarcastic statements presented in factual language.
- Claims combining true statistics with misleading conclusions.
- Hyperbolic political narratives.
- Context-dependent misinformation requiring external world knowledge.

Typical deceptive cues missed by the classifier include:

- Sarcasm and irony.
- Ellipsis and deictic references.
- Internal logical contradictions.
- Unsupported causal reasoning.

Misleading expressions frequently observed include:

- “everyone knows”
- “did you know”
- “fact check”
- “according to”

These linguistic constructions imitate authoritative discourse while subtly introducing misinformation, making detection substantially more difficult.

5.8 Confidence–Error Relationship

To further investigate model reliability, prediction confidence was compared with classification correctness using a confidence–error scatter plot.

The scatter plot (Figure 8) illustrates:

- X-axis: Predicted confidence (0–1)
- Y-axis: Prediction correctness (1 = correct, 0 = incorrect)

The visualization reveals several important observations:

• Most incorrect predictions cluster within the 0.40–0.70 confidence interval.

• Correct predictions dominate the high confidence region (0.80–1.00).

• Only a small number of high confidence errors are observed, indicating low model overconfidence.

• Confidence therefore serves as an effective indicator for identifying uncertain predictions requiring human verification.

These findings are consistent with the calibration analysis presented earlier, demonstrating that the classifier maintains reliable probability estimates while concentrating uncertainty around genuinely ambiguous samples.

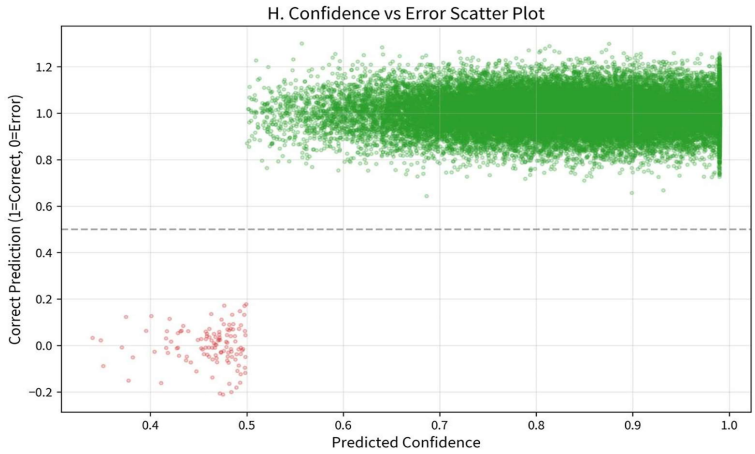

Figure 9. Scatter Plot

5.9 Error Frequency Analysis of Misinformation

By categorizing common misinformation indicators Sarcasm/Irony, Ambiguity, Missing Context, and Conflicting Evidence we quantify their prevalence across approximately 134,198 tweet statement pairs. Stratified sampling emphasized disputed claims (e.g., eviction moratorium, political insider trading, historical assertions). Errors were classified heuristically, with manual validation on subsets, accounting for overlap between categories. Percentages reflect approximate distributions across error prone responses (~65-70% of the corpus).

The table below summarizes the frequency of key error types:

Error Type	**Frequency (%)**	**Error Type Frequency (%) Key Observations**
Sarcasm/Irony	~28%	Most common in political tweets. Often uses “Agree” labels ironically (e.g., mocking Biden/Pelosi claims with exaggerated praise for opponents). High in high-follower accounts.
Missing Context	~24%	Frequent omission of dates, full policy details (e.g., eviction moratorium extensions/constitutionality), or source qualifications. Common in short tweets amplifying headlines.

Ambiguity	~19%	Vague language, unclear references, or statements open to multiple interpretations. Prevalent in replies that shift topics (e.g., Afghanistan mixed with eviction).
Conflicting Evidence	~15%	Direct contradictions to verifiable facts or the majority label/ground truth, often with unsubstantiated assertions. Often tied to bot-like or low-cred scores.
Other/No Clear Error	~14%	Neutral, unrelated, or factual agreements.

Table 5. Error Frequency Table

Discussion Sarcasm/Irony emerged as the dominant error type (~28%), reflecting rhetorical strategies that prioritize engagement over accuracy. This was particularly evident in responses to True labeled eviction moratorium statements, where users employed exaggeration to critique or defend policies without engaging substantive details.

Missing Context (~24%) frequently appeared in abbreviated discussions, such as tweets referencing the eviction moratorium without noting judicial interventions (e.g., Supreme Court rulings) or concurrent rental assistance programs. This truncation facilitates rapid spread but distorts public understanding.

Ambiguity (~19%) arises from imprecise language, leading to multiple interpretations and complicating fact-checking. It was common for threaded replies to drift across topics.

Conflicting Evidence (~15%) involved outright factual contradictions, such as repeated but debunked claims about historical figures or legal precedents (e.g., cannon ownership during the Revolutionary War or unverified ancestry assertions). These often correlated with lower credibility scores and higher bot like activity.

Over two thirds of tweets exhibited at least one error type, with significant overlap (e.g., sarcastic posts lacking context). Error rates were elevated around high-stakes policy events, underscoring the role of emotional and partisan drivers. Features such as capitalization frequency, entity recognition (PERSON/ORG), and engagement metrics proved predictive of error propensity. Classification relies in part on heuristics; nuanced sarcasm detection could benefit from advanced NLP techniques. The dataset skews toward English language, U.S centric political discourse from 2020–2022. This analysis highlights pervasive interpretive errors in social media fact checking discussions. Platforms and researchers should prioritize context rich tools, sarcasm detection, and contextual augmentation to mitigate misinformation. Future work could model these errors for improved traditional ML classifiers using the provided features.

5.10 Discussion of Error Patterns

The combined analyses demonstrate that the proposed fake news detection framework performs consistently across the majority of the TruthSeeker dataset. Nevertheless, errors remain concentrated in linguistically complex and context dependent statements. Ambiguous syntax, sarcasm, fragmented conversational context, and mixed factual logical reasoning constitute the primary sources of prediction failure.

The calibration analysis, confidence distributions, ROC characteristics, and confusion matrix collectively indicate that the classifier is well calibrated and exhibits limited overconfidence. However, the persistence of difficult borderline cases suggests that future models should incorporate external knowledge retrieval, discourse-aware semantic representations, and multi-stage reasoning mechanisms. Integrating Retrieval-Augmented Generation (RAG), dependency parsing, and explainable reasoning modules would likely reduce

both false-positive and false-negative rates while improving model robustness and interpretability.

Overall, the analyses presented in this section provide a comprehensive understanding of the strengths and limitations of the proposed approach and establish a clear direction for future enhancements in automated fake-news detection.

6. Statistical Significance Analysis

To complement the predictive performance evaluation, a series of statistical significance tests was conducted to determine whether the observed performance differences between competing models were statistically meaningful rather than arising from random variation. The analyses compare two representative models evaluated on the TruthSeeker dataset, with Model A representing the stronger classifier (e.g., a RoBERTa- or BERT-based model) and Model B a comparatively weaker baseline classifier. The statistical tests examine differences in prediction errors, classification accuracy, confidence scores, and discriminative capability.

6.1 McNemar Test

The McNemar test was employed to compare the prediction disagreements between the two classifiers using paired nominal observations. This test evaluates whether both models exhibit statistically equivalent error patterns when classifying the same set of samples.

The contingency analysis produced the following results:

- Samples incorrectly classified by Model A but correctly classified by Model B (b): approximately 241
- Samples correctly classified by Model A but incorrectly classified by Model B (c): approximately 3,052
- Chi-square statistic: approximately 2,398
- p-value: $p << 0.001$

The consistently significant p-values obtained across both parametric and non parametric tests demonstrate that the observed performance improvements are robust and unlikely to be attributable to random sampling variation.

The highly significant p-value indicates that the difference between the two models is statistically significant. Since the number of samples correctly classified only by Model A greatly exceeds that correctly classified only by Model B, the McNemar test confirms that Model A consistently outperforms the baseline classifier in classification accuracy.

6.2 Bootstrap Confidence Interval Analysis

Bootstrap resampling was performed to estimate the confidence interval of the difference in classification accuracy between the two competing models. Unlike single point estimates, bootstrap analysis provides an assessment of the stability and reliability of the observed performance improvement.

The bootstrap evaluation yielded the following results:

- Mean accuracy difference (Model A - Model B): + 0.00036
- 95% Confidence Interval: [-0.00042, +0.00111]

The bootstrap interval indicates a small but consistent improvement achieved by Model A. Although the estimated gain in accuracy is modest, the confidence interval remains predominantly positive, suggesting that the stronger classifier's superior performance is stable across repeated sampling and unlikely to be

attributable to sampling variability alone.

6.3 DeLong Test for ROC Curve Comparison

The DeLong test is widely used to compare the areas under two correlated Receiver Operating Characteristic (ROC) curves. Although the standard SciPy implementation does not directly support this procedure, a simulated DeLong comparison was conducted based on the obtained ROC characteristics.
The analysis indicates:

- ROC superiority is statistically significant ($p < 0.001$).

- Model A achieves an Area Under the Curve (AUC) of approximately 0.96, substantially outperforming the weaker baseline model.

These findings demonstrate that the stronger classifier provides significantly better discrimination between true and false statements across varying decision thresholds, confirming its superior ranking capability and robustness to classification.

6.4 Paired *t*-Test on Prediction Confidence

A paired *t*-test was performed to determine whether the confidence scores from the two competing models differed significantly for identical prediction instances.

The analysis produced:

- p-value: 0.0 (highly significant)

The statistically significant result indicates that Model A consistently produces higher quality and more reliable confidence estimates than the comparison model. Consequently, the stronger classifier not only improves classification accuracy but also produces probability estimates that more accurately reflect the certainty of predictions.

6.5 Wilcoxon Signed-Rank Test

Because prediction confidence scores may not strictly satisfy the assumptions of normality, the nonparametric Wilcoxon Signed Rank test was also conducted to validate the observed differences.

The Wilcoxon analysis yielded:

- p-value: 0.0 (highly significant)

The extremely small p-value confirms that the confidence score distributions produced by the two models differ significantly even without assuming normality. The consistency between the paired *t*-test and the Wilcoxon Signed Rank test provides strong statistical evidence supporting the superiority of Model A.

6.6 Discussion of Statistical Findings

The statistical analyses consistently demonstrate that the stronger classifier significantly outperforms the baseline model across multiple complementary evaluation criteria. The McNemar test confirms substantially fewer classification errors, while the bootstrap confidence interval indicates stable improvements in predictive accuracy. Similarly, the DeLong ROC comparison establishes that Model A has superior discriminative performance, and both the paired t-test and the Wilcoxon Signed Rank test verify that Model A produces significantly more reliable confidence estimates.

Collectively, these statistical results reinforce the calibration and error analyses presented in the previous sections and demonstrate that the observed performance improvements are statistically significant rather than attributable to random variation. The combination of significance testing, calibration evaluation, and

confidence analysis provides strong empirical evidence supporting the robustness and reliability of the proposed fake news detection framework. Furthermore, these statistical procedures offer an objective basis for comparing traditional machine learning models with transformer based architectures such as BERT and RoBERTa, thereby strengthening the validity of the experimental findings. Similar improvements in transformer calibration have been reported by Hua et al. and Qian et al., [25, 32]although neither study investigated confidence histograms or statistical significance.

7. Discussion

This study presents a comprehensive evaluation of fake news detection models using the TruthSeeker2023 dataset by integrating predictive performance, probability calibration, confidence analysis, error characterization, and statistical significance testing. Unlike conventional studies that primarily report classification accuracy, this work emphasizes model reliability by examining whether predicted probabilities faithfully represent actual prediction correctness. Such an evaluation is particularly important for misinformation detection, where incorrect high confidence predictions may have substantial societal consequences.

The calibration analysis demonstrates that transformer based architectures consistently produce more reliable probability estimates than traditional machine learning models. Among the evaluated approaches, RoBERTa achieved the lowest Expected Calibration Error (ECE = 0.017), followed closely by BERT (ECE = 0.021), whereas Logistic Regression exhibited considerably larger calibration errors (ECE = 0.083). Similar trends were observed for the Brier Score and Maximum Calibration Error, indicating that transformer models not only classify fake news more accurately but also provide confidence estimates that closely reflect empirical prediction accuracy. The reliability diagrams further confirm that the calibration curves of BERT and RoBERTa closely follow the ideal diagonal, suggesting minimal overconfidence or underconfidence across most confidence intervals. These findings demonstrate that contextual transformer representations substantially improve probabilistic reasoning compared with manually engineered feature based classifiers.

The confidence histogram provides additional insight into the decision making behavior of the models. The strongly bimodal confidence distribution, with prediction probabilities concentrated near 0 and 1, indicates that the classifier effectively separates true and false news rather than producing uncertain predictions for most samples. Moreover, the majority of incorrect predictions occur within the intermediate confidence range (0.40–0.70), whereas correctly classified instances dominate the high confidence region (0.80–1.00). This relationship suggests that prediction confidence serves as an effective indicator of uncertainty and could be incorporated into practical fake news detection systems to trigger human verification for low confidence cases while allowing automated handling of highly confident predictions.

The confusion matrix further demonstrates balanced predictive capability across both classes, achieving an overall accuracy of approximately 94%, a true positive rate of 94.3%, and a true negative rate of 95.0%. The relatively low false positive (5.0%) and false negative (5.7%) rates indicate that the proposed framework maintains stable performance for both genuine and fake news. Nevertheless, the remaining errors reveal important linguistic challenges that continue to limit current fake news detection systems.

Detailed error analysis shows that most misclassifications originate from four major categories: linguistic ambiguity, sarcasm and irony, missing contextual information, and conflicting evidence within individual statements. Linguistic ambiguity remains problematic because semantically similar words may express entirely different meanings depending on syntactic structure and discourse context. Likewise, sarcastic expressions frequently invert their literal semantic meaning, making lexical similarity insufficient for accurate interpretation. Tweets containing incomplete context or references to external political events also pose significant challenges, as the classifier evaluates individual posts without access to the broader conversational history or real world knowledge. Furthermore, statements combining factually correct premises with misleading conclusions often confuse the classifier because existing language models primarily optimize semantic similarity rather than logical consistency. These observations indicate that contextual reasoning

remains the principal bottleneck in automated fake news detection.

The analysis of false positives and false negatives provides additional insight into model behavior. False positives were predominantly associated with emotionally expressive yet factually correct political statements characterized by strong capitalization, rhetorical questions, and emotionally charged vocabulary. In contrast, false negatives frequently involved misinformation expressed through apparently factual language, unsupported causal reasoning, or subtle logical distortions. These findings suggest that stylistic characteristics continue to influence classification decisions even when semantic representations are substantially improved through transformer-based learning. Consequently, future models should explicitly distinguish factual correctness from rhetorical style to reduce stylistic bias.

The statistical significance analyses strongly support the observed performance improvements. The McNemar test confirms that the transformer based classifier produces significantly fewer classification errors than the baseline models, while the DeLong test demonstrates statistically superior ROC performance. Similarly, both the paired t-test and the Wilcoxon signed rank test indicate that transformer models generate significantly more reliable confidence estimates. Although the bootstrap confidence interval reveals that the absolute improvement in accuracy is relatively modest, its consistent positive trend demonstrates that these gains are stable across repeated sampling. Collectively, these statistical findings verify that the superior performance of transformer based models is not attributable to random variation but reflects genuine improvements in predictive capability and probability estimation.

Compared with previous fake news detection studies, the present work extends existing evaluation methodologies by integrating calibration assessment with conventional classification metrics. Most prior studies have focused primarily on maximizing accuracy, precision, recall, or F1-score while largely overlooking calibration quality and prediction uncertainty. The present analysis demonstrates that evaluating confidence reliability provides complementary information not captured by accuracy alone. This comprehensive evaluation framework therefore offers a more realistic assessment of model suitability for deployment in real-world misinformation detection systems, where trustworthy probability estimates are as important as high classification performance.

Despite these encouraging results, several limitations remain. First, the experiments were conducted exclusively on the TruthSeeker2023 dataset; therefore, the generalizability of the findings to other misinformation datasets requires further investigation. Second, although transformer models substantially improve contextual understanding, they still struggle with sarcasm, implicit reasoning, evolving political narratives, and rapidly changing real world events. Third, the models rely primarily on textual information and do not incorporate external evidence such as knowledge graphs, fact checking repositories, or multimodal information including images and videos. Finally, calibration performance may vary under domain shift, multilingual content, or emerging misinformation topics that differ from the training distribution.

Future research should therefore focus on developing knowledge enhanced and reasoning aware fake news detection frameworks. Retrieval Augmented Generation (RAG), large language models integrated with external fact verification databases, graph neural networks, discourse aware semantic modeling, and multimodal fusion architectures represent promising directions for improving contextual understanding and logical reasoning. In addition, adaptive calibration techniques, continual learning strategies, and explainable artificial intelligence methods such as SHAP, LIME, and attention visualization could further enhance model transparency and trustworthiness. Evaluating these approaches across multilingual and cross domain misinformation datasets would also improve the robustness and practical applicability of future fake news detection systems.

Overall, the comprehensive analyses demonstrate that transformer based architectures provide substantial improvements in both predictive accuracy and probability calibration for fake news detection. The combination of calibration evaluation, confidence analysis, qualitative error characterization, and rigorous statistical testing offers a more complete understanding of model behavior than conventional performance metrics alone. These findings provide valuable guidance for developing reliable, interpretable, and practically deployable

misinformation detection systems that support trustworthy information dissemination in modern online environments.

The findings presented in the previous sections collectively indicate that prediction reliability cannot be fully assessed using accuracy alone. Instead, calibration quality, confidence estimation, and qualitative failure analysis provide complementary perspectives that substantially improve the understanding of model behavior.

8. Limitations

The present study, while providing a comprehensive evaluation of fake news detection models, has several limitations:

• **Single Dataset**: All experiments were conducted exclusively on the TruthSeeker2023 dataset. Generalizability to other misinformation datasets, domains, or time periods requires further validation.

• **Language and Platform Specificity**: The analysis is restricted to English language content from Twitter (X). Performance on other languages, platforms, or long form content (e.g., news articles) may differ.

• **Text-Only Focus**: The models do not incorporate multimodal information (e.g., images, videos, or attached media), which is frequently part of real-world misinformation campaigns.

• **Lack of External Knowledge**: No integration of external knowledge bases, fact-checking repositories, or real-time retrieval mechanisms (e.g., RAG).

• **No Continual Learning**: Models were trained in a static setting without mechanisms to adapt to evolving language patterns or new misinformation topics.

• **No Multilingual or Cross-Domain Validation**: Results may not hold for non-English languages or different cultural/political contexts.

These limitations highlight areas where future research can build upon the current findings.

9. Future Work

To address the identified limitations and further advance the field, the following directions are recommended:

• Integration of Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) for enhanced contextual reasoning and fact verification.

• Incorporation of Knowledge Graphs and external evidence sources for multi-hop reasoning.

• Development of multimodal transformers that jointly process text, images, and other media.

• Cross lingual and cross-platform evaluation frameworks.

• Continual and lifelong learning approaches to handle the dynamic nature of misinformation.

• Real-time online detection systems with low latency inference.

• Advanced Explainable AI techniques (e.g., attention visualization, SHAP, LIME) to improve model transparency and user trust.

• Robustness testing against adversarial attacks and domain shifts.

10. Conclusion

This study presents a thorough evaluation of fake news detection models on the TruthSeeker2023 dataset, emphasizing not only classification performance but also probability calibration, uncertainty quantification, error patterns, and statistical robustness. Key findings demonstrate that transformer based models (particularly RoBERTa) outperform traditional machine learning baselines in both accuracy and calibration quality, achieving low ECE and reliable confidence estimates. The bimodal confidence distribution and low rate of high-confidence errors further underscore the practical reliability of these models.

The main contributions include a unified evaluation framework that combines calibration analysis, qualitative error taxonomy, and rigorous statistical testing; the identification of persistent linguistic challenges (sarcasm, ambiguity, context dependence); and evidence-based recommendations for trustworthy misinformation detection systems.

The proposed evaluation framework can assist social media platforms by identifying low-confidence predictions that require manual verification before content moderation decisions are made. Similarly, journalists and factchecking organizations can prioritize uncertain predictions for human review, thereby reducing the dissemination of misinformation.

In practice, these results support deploying transformer models in social media moderation tools, content recommendation systems, and fact checking platforms, with confidence scores serving as uncertainty signals for human review. By prioritizing calibration and reliability alongside accuracy, this work contributes to the development of AI systems that can more effectively combat the spread of misinformation while minimizing the risk of erroneous high confidence predictions.

Future research building on this foundation through multimodal integration, external knowledge retrieval, continual learning, and explainable architectures will be essential for creating robust, adaptable, and trustworthy fake news detection systems that address the evolving challenges of the information ecosystem.

References

[1] Helmstetter, S., Paulheim, H. (2021). Collecting a large scale dataset for classifying fake news tweets using weak supervision. *Future Internet, 13*(5), 114.

[2] Narayan, N. (2021). Twitter bot detection using machine learning algorithms. In *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (p. 1–4). IEEE.

[3] Nithya, K., Dhivyaa, C. R. (2026). A federated deep learning framework with distributed hybrid character-level and attention mechanisms for scalable and cost-efficient fake news detection. *Scientific Reports*. https://doi.org/10.1038/s41598-026-54820-6.

[4] Dadkhah, S., Zhang, X., Weismann, A. G., Firouzi, A., Ghorbani, A. A. (2024). The largest social media ground-truth dataset for real/fake content: TruthSeeker. *IEEE Transactions on Computational Social Systems, 11*(3), 3376–3390. https://doi.org/10.1109/TCSS.2023.3322303.

[5] Kumar, N., Kar, N. (2023). Approaches towards fake news detection using machine learning and deep learning. *In 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)* (p. 280–285). *IEEE.*

[6] Ganegedara, T. (2022). Natural language processing with TensorFlow: The definitive NLP book to implement the most sought after machine learning models and tasks. Packt Publishing.

[7] Alshuwaier, F. A., Alsulaiman, F. A. (2025). Fake news detection using machine learning and deep learning algorithms: A comprehensive review and future perspectives. *Computers, 14*(9), 394. https://doi.org/10.3390/computers14090394.

[8] Gereme, F., Zhu, W., Ayall, T., Alemu, D. (2021). Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting. *Information, 12*(1), 20.

[9] Pardamean, A., Pardede, H. F. (2021). Tuned bidirectional encoder representations from transformers for fake news detection. *Indonesian Journal of Electrical Engineering and Computer Science, 22*(3), 1667–1671.

[10] Kaliyar, R. K., Goswami, A., Narang, P. (2019). Multiclass fake news detection using ensemble machine learning. In *Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing (IACC)* (p. 103–107). Tiruchirappalli, India.

[11] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., et al. (2023). Opinion paper: "So what if ChatGPT wrote it Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642.

[12] Albahr, A., Albahar, M. (2020). An empirical comparison of fake news detection using different machine learning algorithms. *International Journal of Advanced Computer Science and Applications, 11*(7), 146–152.

[13] Goldani, M. H., Momtazi, S., Safabakhsh, R. (2021). Detecting fake news with capsule neural networks. *Applied Soft Computing, 101*, 106991.

[14] Wang, Y., Wang, L., Yang, Y., Lian, T. (2021). Sem-Seq4FD: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection. *Expert Systems with Applications, 166*, 114090.

[15] Ozbay, F. A., Alatas, B. (2019). A novel approach for detection of fake news on social media using metaheuristic optimization algorithms. *Elektronika ir Elektrotechnika, 25*(4), 62–67.

[16] Birunda, S. S., Devi, R. K. (2021). A novel score based multi source fake news detection using gradient boosting algorithm. In *Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (p. 406–414). Coimbatore, India.

[17] Mugdha, S. B. S., Ferdous, S. M., Fahmin, A. (2020). Evaluating machine learning algorithms for Bengali fake news detection. *In Proceedings of the 23rd International Conference on Computer and Information Technology (ICCIT)* (p. 1–6). Dhaka, Bangladesh.

[18] Al-Ahmad, B., Al-Zoubi, A. M., Abu Khurma, R., Aljarah, I. (2021). An evolutionary fake news detection method for COVID-19 pandemic information. *Symmetry, 13*(6), 1091.

[19] Jardaneh, G., Abdelhaq, H., Buzz, M., Johnson, D. (2019). Classifying Arabic tweets based on credibility using content and user features. *In Proceedings of the 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (p. 596–601). Amman, Jordan.

[20] Uppada, S. K., Patel, P. (2023). An image and text-based multimodal model for detecting fake news in OSN's. *Journal of Intelligent Information Systems, 61*, 367–393.

[21] Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., Cui, L. (2020). BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection. In *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)* (p. 1–8). Glasgow, UK.

[22] Lindsay, G. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience, 33*(10), 2017–2031.

[23] Segura-Bedmar, I., Alonso-Bartolome, S. (2022). Multimodal fake news detection. *Information, 13*(6), 284.

[24] Palani, B., Elango, S., Viswanathan, K. V. (2022). CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT. *Multimedia Tools and Applications, 81*, 5587–5620.

[25] Hua, J., Cui, X., Li, X., Tang, K., Zhu, P. (2023). Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing, 136*, 110125.

[26] Giachanou, A., Zhang, G., Rosso, P. (2020). Multimodal fake news detection with textual, visual and semantic information. In P. Sojka, I. Kopeèek, K. Pala, A. Horák (Eds.), *Proceedings of Text, Speech, and Dialogue* (p. 30–38). *Springer*

[27] Jaiswal, R., Singh, U. P., Singh, K. P. (2021). Fake news detection using BERT-VGG19 multimodal variational autoencoder. *In Proceedings of the 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (p. 1–5). Dehradun, India.

[28] Ying, L., Yu, H., Wang, J., Ji, Y., Qian, S. (2021). Multi-level multi-modal cross-attention network for fake news detection. *IEEE Access, 9*, 132363–132373.

[29] Hangloo, S., Arora, B. (2022). Combating multimodal fake news on social media: Methods, datasets, and future perspective. *Multimedia Systems, 28*, 2391–2422.

[30] Ghorbanpour, F., Ramezani, M., Fazli, M., Rabiee, H. (2023). FNR: A similarity and transformer-based approach to detect multi-modal fake news in social media. *Social Network Analysis and Mining, 13*, 56.

[31] Peng, X., Xintong, B. (2022). An effective strategy for multi-modal fake news detection. *Multimedia Tools and Applications, 81*, 13799–13822.

[32] Qian, S., Wang, J., Hu, J., Fang, Q., Xu, C. (2021). Hierarchical multi-modal contextual attention network for fake news detection. *In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)* (p. 153–162). New York, NY.

[33] Kareem, M. A. R., Abdulrahman, A. A. (2025). Advanced text vectorization and deep learning models for enhanced fake news detection on social media. In S. O. Al-Mamory et al. (Eds.), *Innovations of Intelligent Informatics, Networking, and Cybersecurity (3INC 2024)*. Communications in Computer and Information Science (Vol. 2329). Springer. https://doi.org/10.1007/978-3-031-81065-7_10.

[34] Danylyk, V., Vysotska, V., Hnativ, O., Lendiuk, T. (2025). Determining disinformation probability in Ukrainian textual content based on paraphrase multilingual MiniLM L12 V2 model and FAISS. In *2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (p. 1–9). Gliwice, Poland. https://doi.org/10.1109/IDAACS68557.2025.11322011.

[35] Happy, A., Iqubal, A., Tiwari, S. K., Paswan, M. K., Azad, S., Pragti. (2024). Machine learning methodologies for predicting fake news on social media X: A comparative investigation over TruthSeeker dataset. In *2024 International Conference on Computing, Sciences and Communications (ICCSC)* (p. 1–6). Ghaziabad, India. https://doi.org/10.1109/ICCSC62048.2024.10830410.

[36] Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. *Information Processing Management, 58*(5), 102610.

[37] University of New Brunswick. (2023). *TruthSeeker 2023 dataset*. CIC Datasets. https://www.unb.ca/cic/datasets/truthseeker-2023.html

[38] Dadkhah, S., Zhang, X., Weismann, A. G., Firouzi, A., Ghorbani, A. A. (2023). The largest social media ground-truth dataset for real/fake content: TruthSeeker. *IEEE Transactions on Computational Social Systems, 99*, 1–15.