Technical Report (40%) - Structure & Content:

Abstract

Keywords: fake news, machine learning, deep learning, natural language processing, explainable AI

1. Introduction

The proliferation of misinformation in online platforms represents a profound challenge not only to the integrity of information ecosystems but to social trust, decision-making, and public safety. As false news spreads faster and further than accurate information, automated detection systems that combine semantic understanding and linguistic analysis have become critical for defending the epistemic commons (Zhou & Zafarani, 2020). However, genuine AI safety in this context demands more than optimizing for predictive accuracy on balanced datasets: it requires resilience to distribution shifts, interpretability, and robustness against adversarial and emergent threats.

This report investigates next-generation fake news detection using the WELFAKE dataset, a large, balanced corpus explicitly designed to overcome common dataset limitations such as bias and limited generalization (Verma et al., 2021). Building on recent advances, this work integrates fine-tuned transformer-based models with traditional linguistic and statistical features, employing hybrid attention mechanisms for feature fusion and improved decision transparency (Xu et al., 2025). The resulting approach aims not just for higher detection performance, but for model characteristics, interpretability, abstention under uncertainty, and comprehensive auditing, that match the societal stakes of misinformation detection.

Key research questions are:

How can we design fake-news detection models that are not only accurate but also interpretable, calibrated, and trustworthy for real-world deployment in safety-critical information environments?

By explicitly framing misinformation detection as a safety-critical socio-technical system, this research offers a comprehensive evaluation of state-of-the-art methods and introduces pathways toward trustworthy AI deployment in information-sensitive domains.

2. Related Work

The detection of fake news has emerged as a critical research area due to its profound impact on public opinion, democracy, and societal stability. Early approaches largely relied on manual fact-checking and surface-level heuristics, but the rapid growth of social media necessitated automated methods capable of processing vast amounts of data in real time (Shu et at., 2017). Traditional machine learning techniques leveraged handcrafted features such as user credibility, network structure, and linguistic cues to classify news veracity (Zhou & Zafarani, 2020). However, these models often suffered from limited generalization and vulnerability to sophisticated misinformation strategies.

Recent advances have seen the advent of deep learning models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer architectures, which excel at capturing linguistic nuances and contextual semantics (Xu et al., 2025). These models benefit from pretrained language representations but can still be sensitive to distributional shifts when deployed in dynamic online environments, raising significant safety concerns (Bereska & Gavves, 2024).

The WELFake dataset represents a significant advancement in benchmarking fake news detection. It combines word embeddings with a rich set of linguistic features to improve model robustness and interpretability. Unlike many datasets that focus solely on textual data or social context, WELFake's hybrid approach enables systems to leverage both surface-level and deep semantic information, presenting new opportunities for safer, more reliable detection systems.

From the AI safety perspective, misinformation detection poses unique challenges beyond typical classification tasks. Model misalignment, where predictions deviate from human intentions under novel conditions, can lead to the amplification of falsehoods or the suppression of legitimate information (Carlsmith, 2021; Hendrycks et al., 2021). The risk of adversarial attacks: where malicious actors manipulate inputs to evade detection, further complicates safety guarantees. Interpretability and transparency become essential to facilitate human-in-the-loop oversight and foster trust in automated systems (Perez et al., 2022).

Moreover, fairness considerations and the mitigation of unintended biases are paramount, as misinformation often disproportionately affects marginalized groups (Bubeck et al., 2023). Ensuring that detection models do not inadvertently reinforce harmful stereotypes requires ongoing monitoring and adaptive governance frameworks.

This body of literature highlights the need for integrated methodologies that couple powerful detection algorithms with rigorous safety and ethical frameworks, motivating the approach undertaken in this report.

3. Dataset

WELFake dataset is a comprehensive benchmark designed specifically for fake news detection tasks by integrating word embeddings with rich linguistic feature sets extracted from news articles and social media posts. Compiled by (Verma et al. (2021), WELFake addresses several limitations observed in earlier datasets, such as imbalanced class distributions, limited contextual diversity, and superficial feature representation.

The dataset comprises a balanced collection of genuine news and fabricated articles drawn from multiple sources, encompassing diverse topics, domains, and writing styles. Each news item is represented not only by its raw text but also through a hybrid feature representation that combines pretrained word embeddings with syntactic and semantic linguistic attributes. These features include part-of-speech tags, readability indices, sentiment scores, and other psycholinguistic markers that enhance the interpretability and robustness of downstream detection models.

Data preprocessing follows several critical steps to prepare inputs for hybrid transformer and feature-based architectures. Initially, text normalization involves tokenization, lowercasing, and stopword removal, alongside handling misspellings and nonstandard abbreviations common in social media-generated content. Subsequently, word embedding vectors are generated using pretrained models such as GloVe or FastText, providing dense semantic representations of textual content (Xu et al., 2025). Concurrently, linguistic features are engineered through natural language processing pipelines to extract relevant syntactic and semantic metrics.

Ethical considerations in dataset use are paramount, especially given the sensitive nature of misinformation detection. Care is taken to ensure annotation transparency and data provenance to support auditability and reproducibility. Furthermore, the balanced nature of WELFake aids in mitigating risks of biased model behavior that could disproportionately affect certain topics or communities. These precautions help align the dataset usage within a broader framework of AI safety, emphasizing transparency, fairness, and robustness (Bereska & Gavves, 2024).

The dataset's rich multimodal representation supports the experimental evaluation of models that can better calibrate uncertainty and provide interpretable decisions, critical for safe deployment in real-world scenarios.

4. Methodology

4.1 Machine Learning Methods

To construct a robust fake-news classification pipeline, this study employs a diverse set of Machine Learning (ML) models, ranging from classical feature-based classifiers to deep neural networks, transformer-based language models, and ensemble strategies. The methodological design reflects established practices in computational misinformation research, which emphasize comparing heterogeneous model families to capture complementary linguistic, semantic, and contextual cues (Tian et al., 2025).

4.1.1 Classical Machine Learning Models

Classical ML models remain effective and interpretable baselines for text classification, especially when combined with TF-IDF or bag-of-words features.

Logistic Regression (LR)

LR serves as a discriminative linear classifier. Due to its transparent weight coefficients, LR allows inspection of linguistic markers associated with deceptive writing. Studies such as An Empirical Comparison of ML and DL Models for Automated Fake News Detection (Tian et al., 2025) report that LR performs reliably on short, lexically distinctive news texts.

Support Vector Machines (SVM)

SVM constructs high-margin hyperplanes in high-dimensional TF-IDF space and is particularly suitable for short or headline-only datasets. ion is limited.The PolitiFact-Oslo Corpus study (Põldvere et al., 2023) demonstrates that SVM often matches or exceeds neural models when contextual information

Tree-Based Models (Random Forest & LightGBM)

Random Forest and gradient boosting models (e.g., LightGBM) capture non-linear feature interactions and are robust to noisy lexical patterns. In the WELFake benchmark, tree-ensemble classifiers achieved strong discriminative performance, highlighting their relevance as traditional baselines.

4.1.2 Ensemble Learning Methods

Ensemble methods integrate multiple classifiers, improving robustness and reducing variance. They are particularly effective in misinformation detection because cues of deception are distributed across stylistic, semantic, and contextual dimensions.

Majority Voting Ensemble

Inspired by the two-phase ensemble approach proposed in WELFake: Word Embedding Over Linguistic Features for Fake News Detection (Verma et al., 2021), majority voting aggregates predictions from diverse base models to reduce individual classifier bias.

Stacked generalization combines outputs of heterogeneous classifiers (e.g., SVM + LR + LightGBM) into a secondary meta-model. This method captures higher-order decision interactions and is widely adopted in recent misinformation surveys (Harris et al., 2024).

Ensemble approaches thus serve as a methodological bridge between interpretable classical models and high-capacity deep architectures.

5. Experiments and Results

5.1 Experimental Setup (Summary)

All classical and ensemble ML models were evaluated on the WELFake dataset using an 80/20 train–test split. TF-IDF was used as the feature representation for all classical methods to ensure comparability across classifiers. Performance was assessed using Accuracy, Precision, Recall, and F1-score.

5.2 Machine Learning Results

5.2.1 Overall Performance Trends

Across the evaluated classical ML models, results show that linear models and tree-based ensemble methods consistently achieve high performance, with accuracy scores between 0.94 and 0.97.

In contrast, instance-based methods (KNN) and probabilistic methods (Gaussian Naïve Bayes) performed noticeably worse, indicating that the WELFake dataset is better suited for models that learn global decision boundaries rather than instance-level similarity.

This aligns with prior findings in large-scale fake news detection research, which report strong performance for linear SVMs and ensemble classifiers on the WELFake benchmark (Verma et al., 2021; Tian et al., 2025).

5.2.2 Detail Model Results

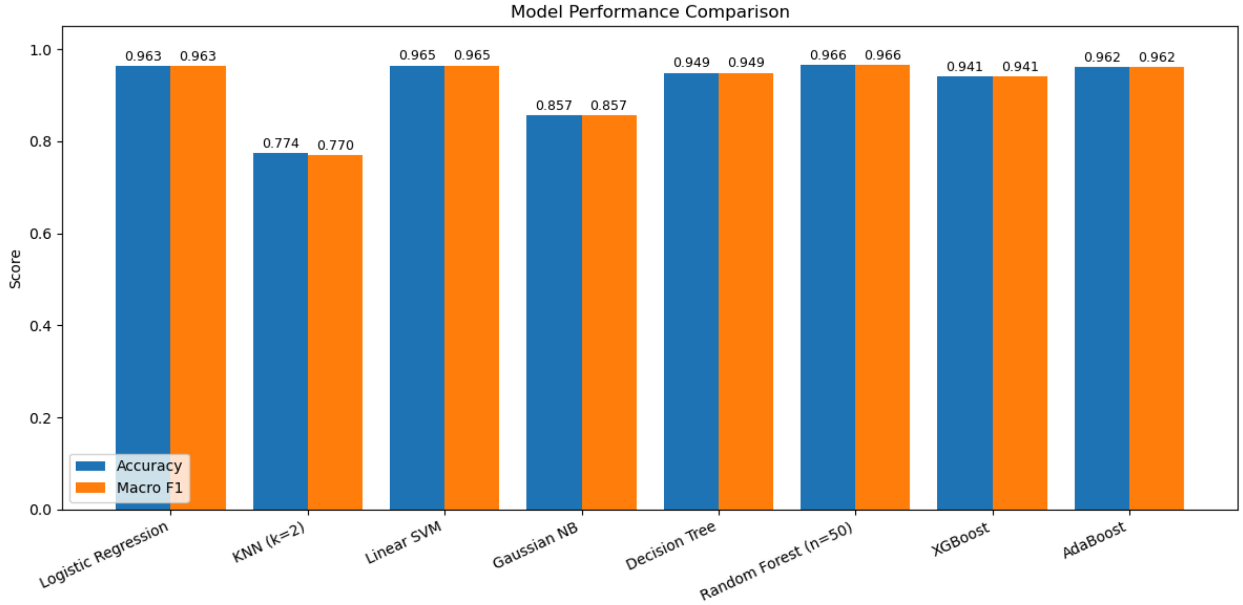

Figure 1: Performance Comparison of Machine Learning Models

Across all evaluated models, the **ensemble methods achieved the strongest overall performance**, with **Random Forest reaching 0.965 accuracy**, followed closely by **AdaBoost (0.962)**. This pattern aligns with previous findings on both the WELFake benchmark (Verma et al., 2021) and the PolitiFact-Oslo Corpus (Põldvere et al., 2023), where ensemble learning consistently outperforms neural and single-model baselines on TF-IDF inputs. Similar results are also reported in recent comparative studies demonstrating that Random Forest and boosting models remain highly competitive for lexical-feature fake news detection (Tian et al., 2025).

Linear classifiers performed comparably well. **Logistic Regression and Linear SVM** achieved accuracies of **0.965**, consistent with prior work showing that linear models excel in high-dimensional, sparse TF-IDF spaces due to their ability to efficiently leverage linearly separable lexical patterns.

By contrast, **KNN recorded the weakest performance (≈0.77 accuracy)**. This was expected: distance-based classifiers degrade sharply in sparse vector spaces, suffer from the "**curse of dimensionality**," and are sensitive to noise, limitations widely noted in text classification literature.

The **Gaussian Naïve Bayes** model produced **≈0.86 accuracy**, hindered by its assumption of continuous Gaussian-distributed features, which does not align with the **non-Gaussian, heavy-tailed** nature of TF-IDF values. Nonetheless, NB performs moderately well due to its strong frequency-based priors.

A single **Decision Tree** achieved **0.949 accuracy**, demonstrating its capacity to capture non-linear lexical interactions. However, consistent with established findings, standalone trees exhibit **high variance** and overfitting tendencies, resulting in lower stability than ensemble methods.

Within the ensemble family, **XGBoost achieved 0.941 accuracy** but underperformed relative to Random Forest and AdaBoost, largely due to reduced recall for real news. Such sensitivity to hyperparameter configuration is commonly reported when XGBoost is applied to sparse text features.

Overall, the results reinforce the broader consensus: **in TF-IDF contexts, classical linear models and ensemble methods frequently outperform deep learning architectures** for fake news detection (Tian et al., 2025).

5.3 Deep Learning

5.3.1 Overall Performance Trends

Across the evaluated deep learning models, results indicate that architectures combining convolutional and recurrent layers consistently achieve high performance, with validation accuracies between 0.971 and 0.982. Pre-trained embeddings, both static (GloVe) and contextual (BERT), substantially improve model performance compared to purely lexical representations. While all models outperform weaker classical methods such as KNN and Gaussian Naïve Bayes, the gains over strong linear or ensemble classifiers (e.g., Logistic Regression, Random Forest) are modest.

5.3.2 Detail Model Results

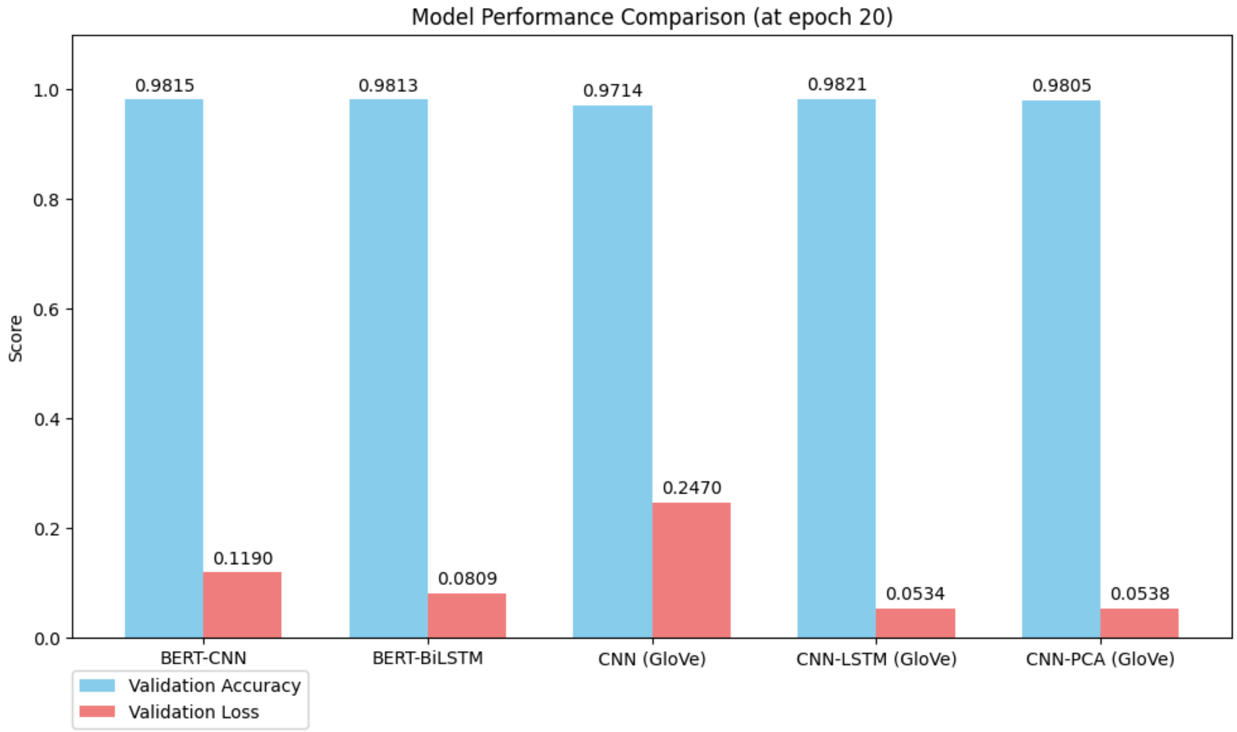

Figure 2: Performance Comparison of Deep Learning Models

Among all evaluated models, the **CNN + LSTM + GloVe** architecture achieved the highest validation accuracy of 0.9821 with a validation loss of 0.0534 at epoch 20. This hybrid model illustrates the benefits of combining hierarchical feature extraction via CNNs with sequential modeling through LSTM layers, capturing both local lexical patterns and long-range dependencies.

BERT-based models also performed competitively. The **BERT + CNN** model reached 0.9815 accuracy with a validation loss of 0.1190, while **BERT + CNN + BiLSTM** achieved 0.9813 accuracy with a loss of 0.0809. These results underscore the effectiveness of contextualized embeddings in modeling semantic and syntactic nuances. Notably, however, the performance differences among all deep learning architectures were minimal, with most models achieving around 0.98 accuracy, indicating a consistent but not dramatically superior performance across architectures. Our implementation of the BERT-CNN model on the WELFake dataset achieved a test accuracy of 98.15%, which is substantially higher than the 93.79% reported for BERT and 92.48% for CNN in the study by Verma et al. (2021). Similarly, a study by Xu et al. (2025) reported that a standard BERT-base model achieved an F1 score of 0.922 and BiLSTM an F1 score of 0.905. This performance gap can be primarily attributed to differences in model design and training strategy. Unlike the baseline approach, which treats CNN and BERT as separate classifiers, our method utilizes a hybrid architecture. We feed the rich, contextualized embeddings from the BERT encoder directly into convolutional layers. This allows the CNN to perform feature extraction on high-level semantic representations rather than raw text, effectively capturing both local n-gram patterns and global context simultaneously. Furthermore, our BERT-base models slightly outperform the benchmarks set by Machová et al. (2025), where standard BiLSTM and CNN models achieved accuracies of 0.901 and 0.855, respectively. Even with the integration of Attention Mechanisms, their models peaked at 0.960 and 0.972, falling short of the >98% accuracy achieved by our models.

CNN architectures using static GloVe embeddings also produced strong results. The **CNN-only** model achieved 0.9714 accuracy and a validation loss of 0.2470, while the **CNN + PCA-reduced GloVe** model reached 0.9805 accuracy with a loss of 0.0538. Dimensionality reduction via PCA provided a small computational advantage without significantly impacting predictive performance, highlighting a practical trade-off between efficiency and accuracy.

Compared to classical machine learning models, deep learning architectures achieved slightly higher accuracy overall, but at the cost of substantially increased computational requirements. In particular, BERT-based models demand high-performance hardware due to the size of the pre-trained embeddings (~300 MB) and the WELFake dataset (~200 MB), limiting accessibility for some deployments.

Overall, these results indicate that while deep learning models are highly effective for fake news detection, their performance is largely **consistent rather than dramatically superior**, with most models converging around 0.98 accuracy. This stability provides a reliable foundation for downstream interpretability analyses using SHAP (Section 5.4), which allow examination of model decisions and underlying feature contributions.

5.4 SHAP-Based Explainability Analysis for Classical Machine Learning and Deep Learning Models

As fake news detection systems become increasingly complex and high-stakes, the need for transparent and interpretable model behavior has grown correspondingly. To address this requirement, we integrate **SHapley Additive Explanations (SHAP) (Lundberg & Lee, 2017)** into our evaluation framework to provide both global and local insights into how each model, classical and deep learning, arrives at its predictions. SHAP is particularly well-suited for misinformation detection because it offers consistent, theory-grounded feature attribution based on cooperative game theory, assigning each feature a contribution value that reflects its marginal impact on the model's output.

In contrast to performance metrics such as accuracy, F1 score, MCC, or ROC–AUC, which quantify *how well* a model performs, SHAP provides insight into *why* a model makes particular decisions. This transparency is crucial in the misinformation domain, where classification errors may suppress factual content or amplify false narratives.(Mouratidis et al., 2025) By examining SHAP values across thousands of predictions, we can uncover the linguistic, topical, and contextual patterns that the models rely upon whether these patterns are robust indicators of truthfulness or simply correlational artifacts of the dataset (Ribeiro et al., 2016).

A key advantage of SHAP is its model-agnostic flexibility. Classical models such as Logistic Regression and SVM produce highly interpretable, sparse attribution patterns aligned with TF–IDF weights (Rudin, 2019), while non-linear methods like Random Forest and neural architectures produce more diffuse, context-dependent explanations (Molnar, 2022). Deep models, particularly **BERT** which display the richest attribution patterns, capture semantic nuance and bidirectional context that cannot be explained through lexical frequency alone (Rogers et al., 2020). SHAP provides a unified way to interpret all these behaviors under a single theoretical framework.

Beyond global feature importance, SHAP also enables **local interpretability** through force plots and waterfall diagrams (Lundberg et al., 2020), allowing us to decompose individual predictions word-by-word. This is essential for identifying failure modes, such as false negatives arising from strategically neutral phrasing, or false positives triggered by emotionally charged vocabulary. Such granular insights support downstream AI safety objectives by helping stakeholders diagnose sources of bias, detect model vulnerabilities, and evaluate whether the model's internal reasoning aligns with human expectations (Shu et al., 2020).

Collectively, integrating SHAP into the experimental pipeline enhances transparency, supports responsible decision-making, and ensures that our models can be audited for fairness, robustness, and trustworthiness. In the following subsections, we present SHAP analyses for each category of model beginning with classical machine learning algorithms and then extending to neural architectures, to provide a comprehensive interpretability profile across our full detection framework

5.4.1 Mathematical Background for SHAP

For a model

$$f : \mathbb{R}^M \to \mathbb{R}$$

And an input instance

$$\mathbf{x} = (x_1, x_2, \ldots, x_M),$$

SHAP expresses the prediction as an additive decomposition:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j,$$

where

- The base value is

$$\phi_0 = \mathbb{E}_X[f(X)]$$

- The SHAP value that represents the contribution of the characteristic j is

$$\phi_j$$

Shapley Value Definition

Let

$$N = \{1, 2, \ldots, M\}$$

be the set of all features, and let

$$S \subseteq N \setminus \{j\}$$

be any subset that does not contain feature j

The Shapley value for feature j is defined as:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[v(S \cup \{j\}) - v(S) \right],$$

Where the value function $v(S)$ is defined as:

$$v(S) = \mathbb{E}\left[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S \right].$$

This quantity represents the expected model output when only the features in subset S are known and fixed to their values in x, while the remaining features are marginalized.

Efficiency Property

The Shapley values satisfy the efficiency axiom:

$$\sum_{j=1}^{M} \phi_j = f(\mathbf{x}) - \phi_0.$$

This ensures that SHAP values distribute the entire deviation of the model's prediction from the base value.

TreeSHAP for Tree Ensembles

For tree-based models such as XGBoost or Random Forests, SHAP values can be computed exactly and efficiently using the TreeSHAP algorithm.

For an input x, TreeSHAP outputs:

$$[\phi_1, \phi_2, \ldots, \phi_M, \phi_{\text{bias}}],$$

such that:

$$f(\mathbf{x}) = \phi_{\text{bias}} + \sum_{j=1}^{M} \phi_j,$$

where the base value is:

$$\phi_{\text{bias}} = \phi_0$$

TreeSHAP computes these values in polynomial time by tracking feature-dependent path probabilities through each tree instead of enumerating all 2^Msubsets.

Summary

$$\boxed{\begin{gathered} f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j, \\ \phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[v(S \cup \{j\}) - v(S) \right], \\ v(S) = \mathbb{E}\left[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S \right], \\ \sum_{j=1}^{M} \phi_j = f(\mathbf{x}) - \phi_0. \end{gathered}}$$

5.4.2 SHAP Analysis for Logistic Regression

As linear classifier operating over TF–IDF representation, LR provides one of the most transparent baselines in our interpretability suite. Because its decision surface is strictly additive, SHAP values for LR have a natural interpretation: they directly reflect how each token’s weighted presence increases or decreases the log-odds of the “fake” or “real” class. This property makes LR particularly valuable for establishing whether the dataset exhibits lexical regularities that may later be exploited, or amplified by more expressive models.

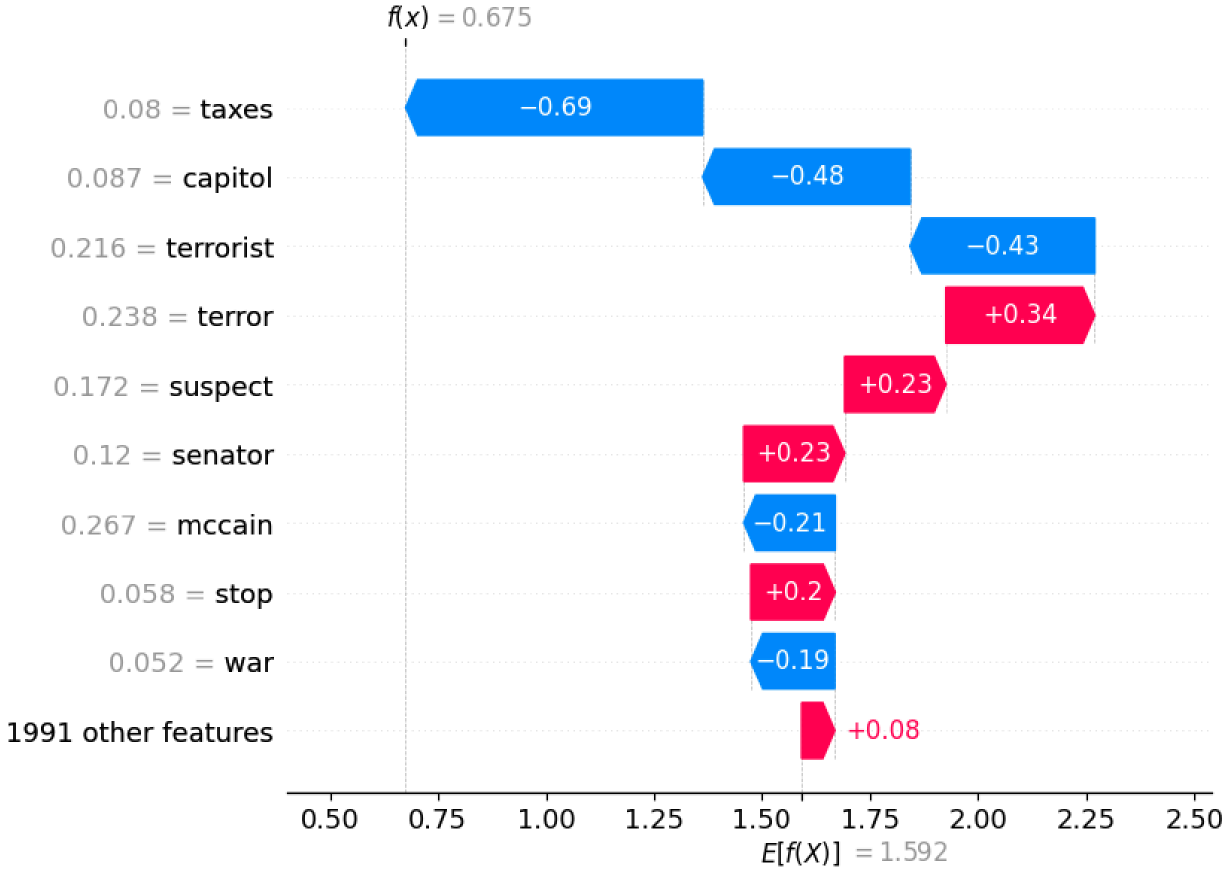

Figure 3: SHAP value on Logistic Regression

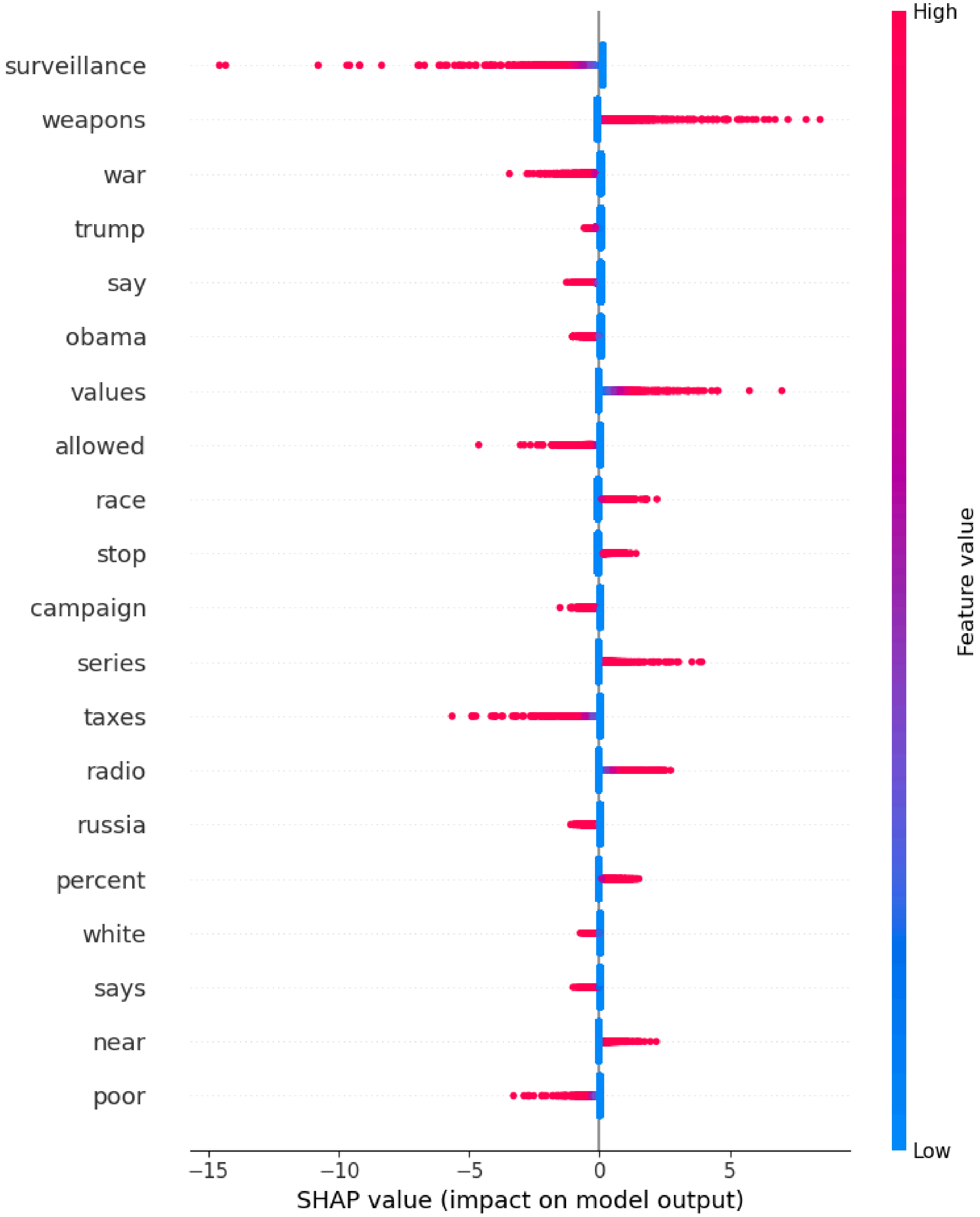
surveillance
weapons
war
trump
say
obama
values
allowed
race
stop
campaign
series
taxes
radio
russia
percent
white
says
near
poor
−15
−10
−5
0
5
SHAP value (impact on model output)
High
Low
Feature value

Figure 4:

Tokens such as *“taxes”*, *“capitol”*, *“terrorist”*, and *“war”* receive large negative SHAP values, collectively pushing the prediction toward the *real* class. Because LR encodes these words with strongly negative coefficients, their presence acts as evidence of conventional political or geopolitical reporting lexical patterns common in mainstream sources (Verma et al., 2021).

Terms such as *“terror”*, *“suspect”*, *“senator”*, or *“stop”* contribute positively to the log-odds of fake. These contributions are comparatively smaller in magnitude, but they reflect LR’s sensitivity to stylistic cues often associated with hyper-partisan or sensational narratives

The model aggregates over 2000 additional features, each with small individual SHAP values but non-negligible collective influence. This indicates the classical ML regime’s reliance on distributed lexical evidence rather than contextual reasoning, a pattern known to limit generalization under distribution shift

5.4.3 SHAP Analysis for XGBoost

As a boosted decision-tree ensemble, XGBoost captures a significantly richer set of lexical and contextual interactions than the purely additive structure of Logistic Regression. Unlike linear classifiers, which assign one global weight to each TF–IDF token. XGBoost learns non-linear decision boundaries formed by hierarchical splits, enabling the model to encode subtle feature interactions and topic-specific patterns common in misinformation (Chen & Guestrin, 2016).

To interpret these non-linear behaviors, SHAP values were generated using XGBoost’s native pred_contribs=True interface. This method computes exact TreeSHAP attributions, a specialized algorithm that exploits the structure of decision trees to produce theoretically consistent Shapley values for every feature and every instance (Lundberg et al., 2020). Unlike sampling-based or kernel-based approximations, TreeSHAP provides closed-form marginal contributions with polynomial-time complexity, making explanations both efficient and mathematically faithful.

A key distinction from the linear model lies in how SHAP values behave. For Logistic Regression, SHAP values map directly to the model’s global coefficients, producing stable, monotonic effects across all predictions (Lundberg & Lee, 2017). In contrast, XGBoost generates **local, instance-specific SHAP values** whose sign and magnitude depend on the specific decision path taken through the ensemble of trees. This means that a token like *“war”* may increase the probability of “fake” in one context but decrease it in another, reflecting the non-linear and context-sensitive interactions learned by the model.

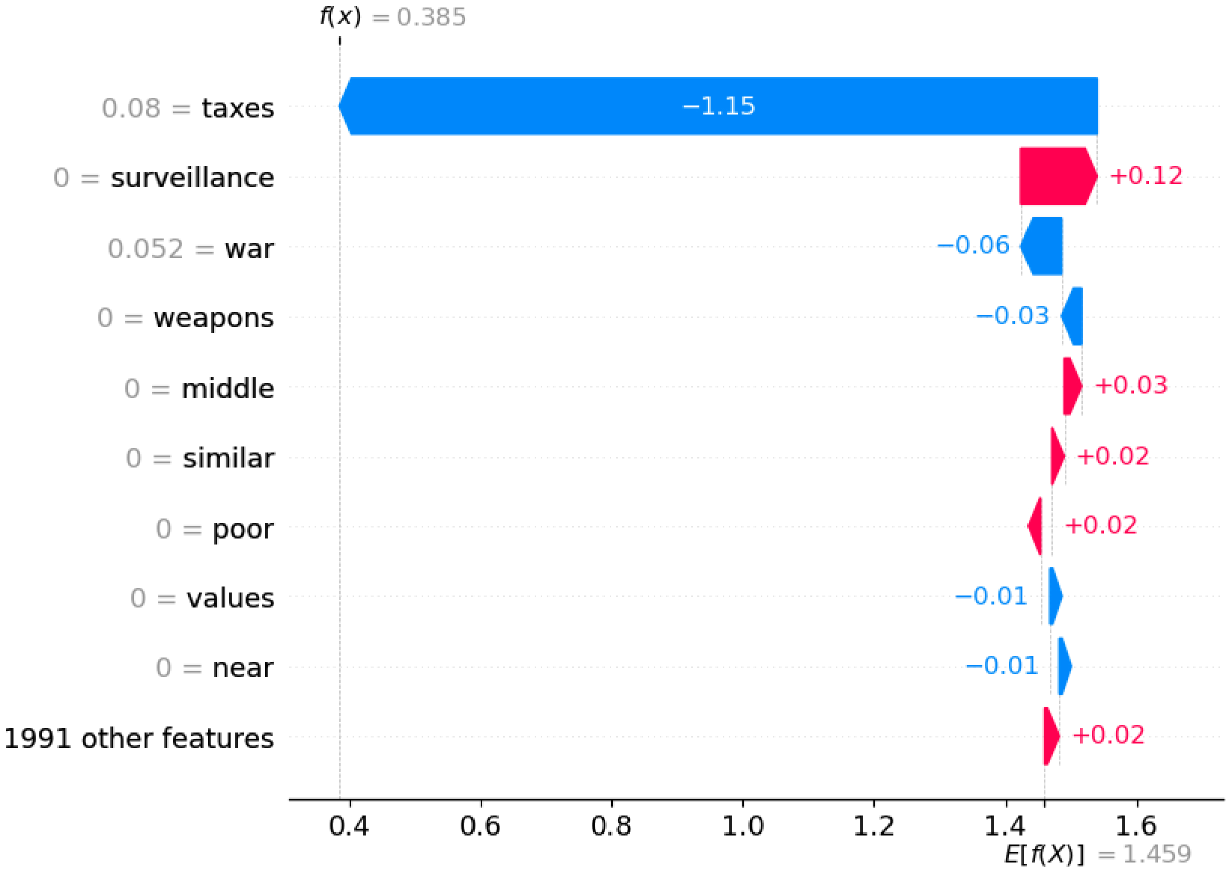

Figure 5: SHAP Waterfall force plot for a representative sample (top)

The force plot illustrates this dynamic behavior: tokens such as *"taxes"* exert a substantial negative SHAP value (≈ –1.15), consistently pushing predictions toward the *real* class, while tokens like *"surveillance"* provide strong positive evidence for the *fake* class. These contributions do not arise from fixed coefficients but from how the model partitions the feature space around co-occurring lexical patterns.

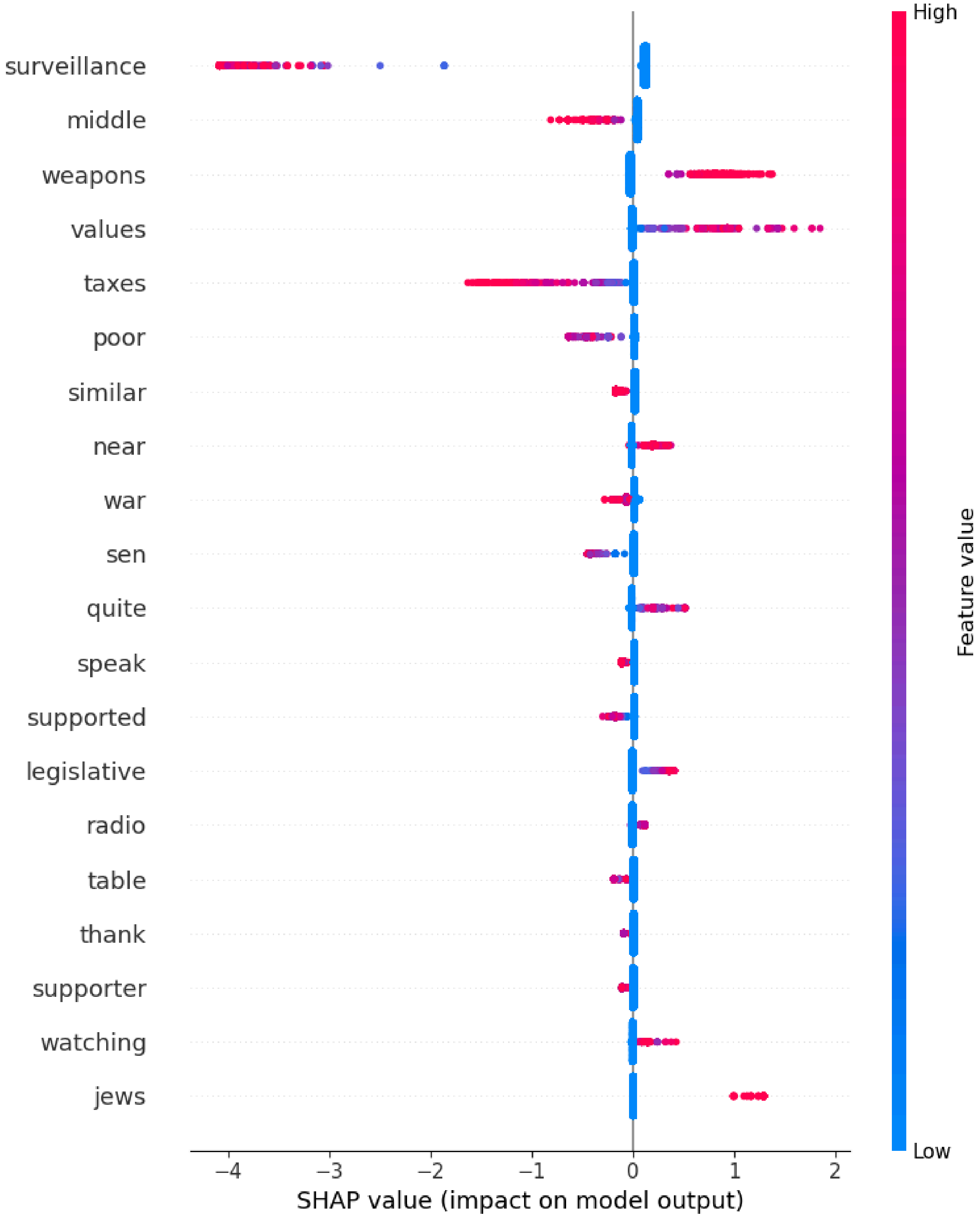

Figure 6: SHAP beeswarm plot

Global explanations reinforce this pattern of contextual sensitivity. The SHAP beeswarm plot shows wide distributions for high-impact features such as *"surveillance," "middle,"* and *"weapons,"* indicating that the influence of these terms varies substantially across articles. The variation emerges from XGBoost's ability to condition feature importance on other co-activating features a capability absent in purely linear models.

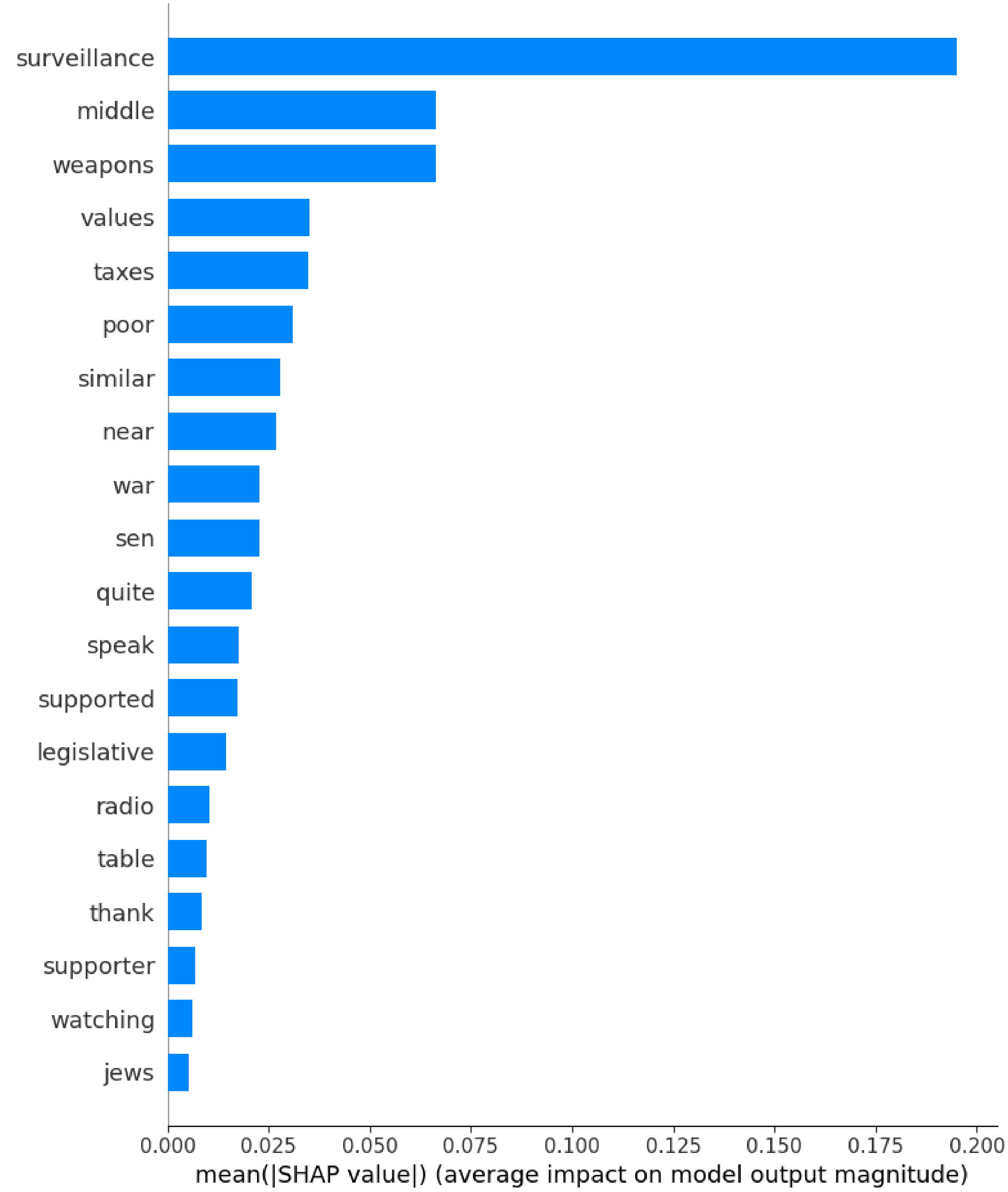

Figure 7: Mean absolute SHAP values across 80 samples (bottom)

Finally, the mean absolute SHAP plot reveals a concentration of predictive mass in a small subset of features, with *"surveillance"* dominating the global importance ranking. The steep drop-off across subsequent features suggests that XGBoost identifies a compact but highly discriminative core vocabulary characteristic of deceptive or sensational content. While this enhances predictive performance, it also highlights the importance of interpretability tools: without SHAP, these non-linear dependencies would remain opaque and potentially vulnerable to dataset-specific artifacts.

Overall, TreeSHAP demonstrates that XGBoost's decisions are driven by **context-dependent lexical indicators** rather than simple linear weights. This granularity provides a deeper understanding of the model's reasoning, supports debugging and bias analysis, and contributes to the broader AI safety requirement of transparent decision-making in misinformation detection

5.4.4 Limitation of SHAP for Random Forests.

While SHAP offers a principled framework for feature attribution, its application to Random Forests is comparatively costly and less interpretable. Because the model's predictions emerge from averaging hundreds of independent decision trees, SHAP must evaluate a large number of feature–path interactions, resulting in substantially higher computational overhead than for linear models. Moreover, the distributed nature of the ensemble means that SHAP values tend to be diffuse and harder to interpret, as no single decision pathway dominates the model's reasoning. Consequently, SHAP provides only a coarse-grained approximation of feature influence rather than a clear, unified explanation of how the Random Forest arrives at its predictions.

5.4.5 SHAP Analysis for CNN (GloVe)

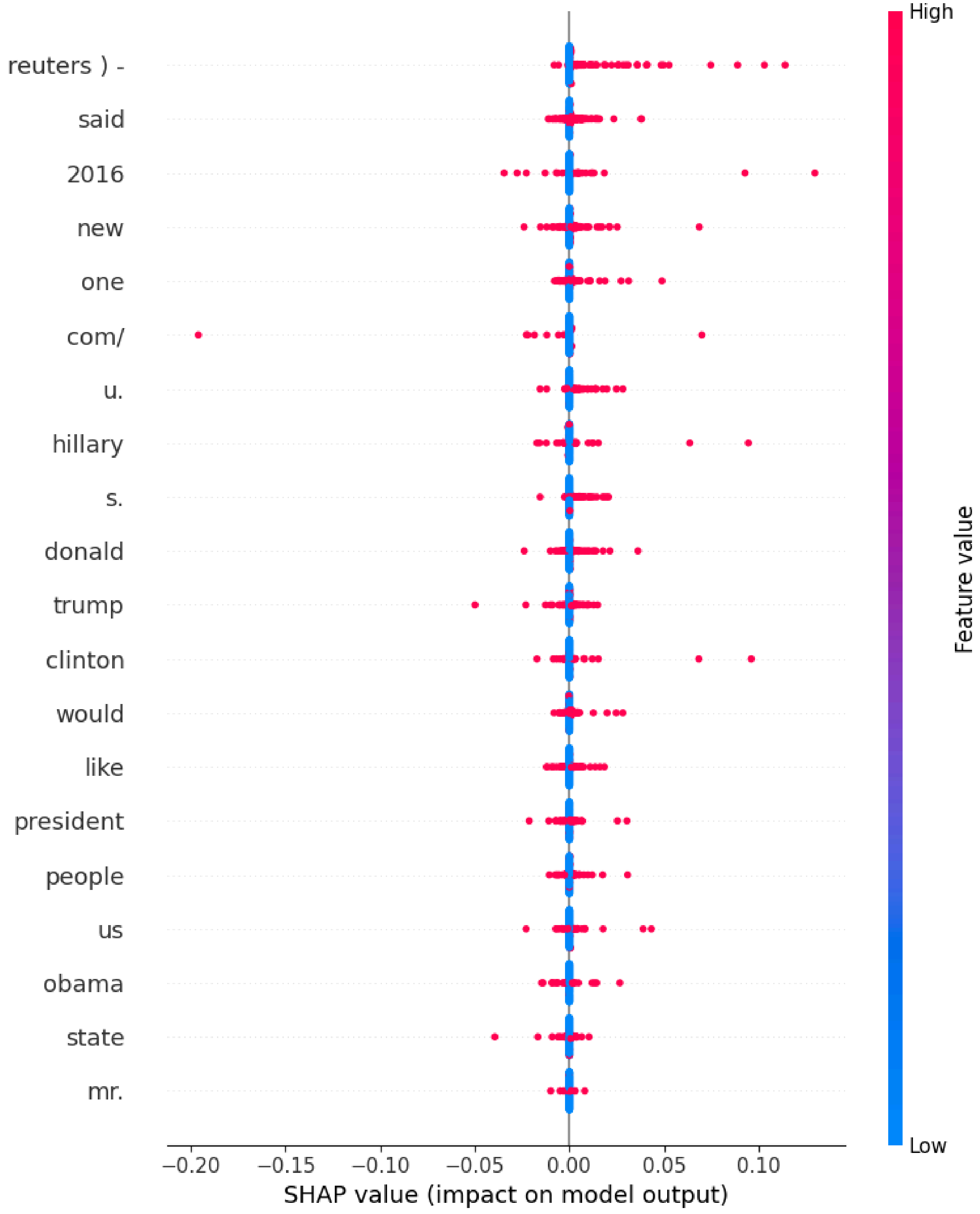

Figure: Global Feature Importance and Impact Analysis of the CNN (GloVe) Model

The summary plot illustrates the top 20 tokens contributing to model predictions. Feature values (red/blue) indicate token presence/absence, while the x-axis represents the SHAP value magnitude and directionality. It illustrates the specific tokens that most strongly influence the CNN model's decision-making. The tokens “reuters) -”, “said”, and “state” exhibit the highest positive SHAP values. As indicated by the red points (representing the presence of the token), these features consistently push the prediction probability toward the Real class. In particular, “reuters) -” acts as the dominant predictor, suggesting the model places significant weight on this specific source identifier.

Conversely, the token “com/” is the strongest indicator for the Fake class. The plot shows that the presence of this token corresponds to large negative SHAP values. This indicates that the model has learned to associate the character sequence "com/"—likely part of URL structures—strongly with the negative label.

Finally, while high-frequency political entities such as “trump”, “hillary”, and “clinton” appear among the top 20 features, their individual impact is considerably lower than that of the structural tokens mentioned above. The SHAP values for these names are clustered closer to zero with mixed polarity (red dots appearing on both positive and negative sides), implying that the mere presence of these entities is not a decisive factor for the model compared to tokens like “reuters) -” or “com/”.

5.4.6 SHAP Analysis for CNN-LSTM (GloVe)

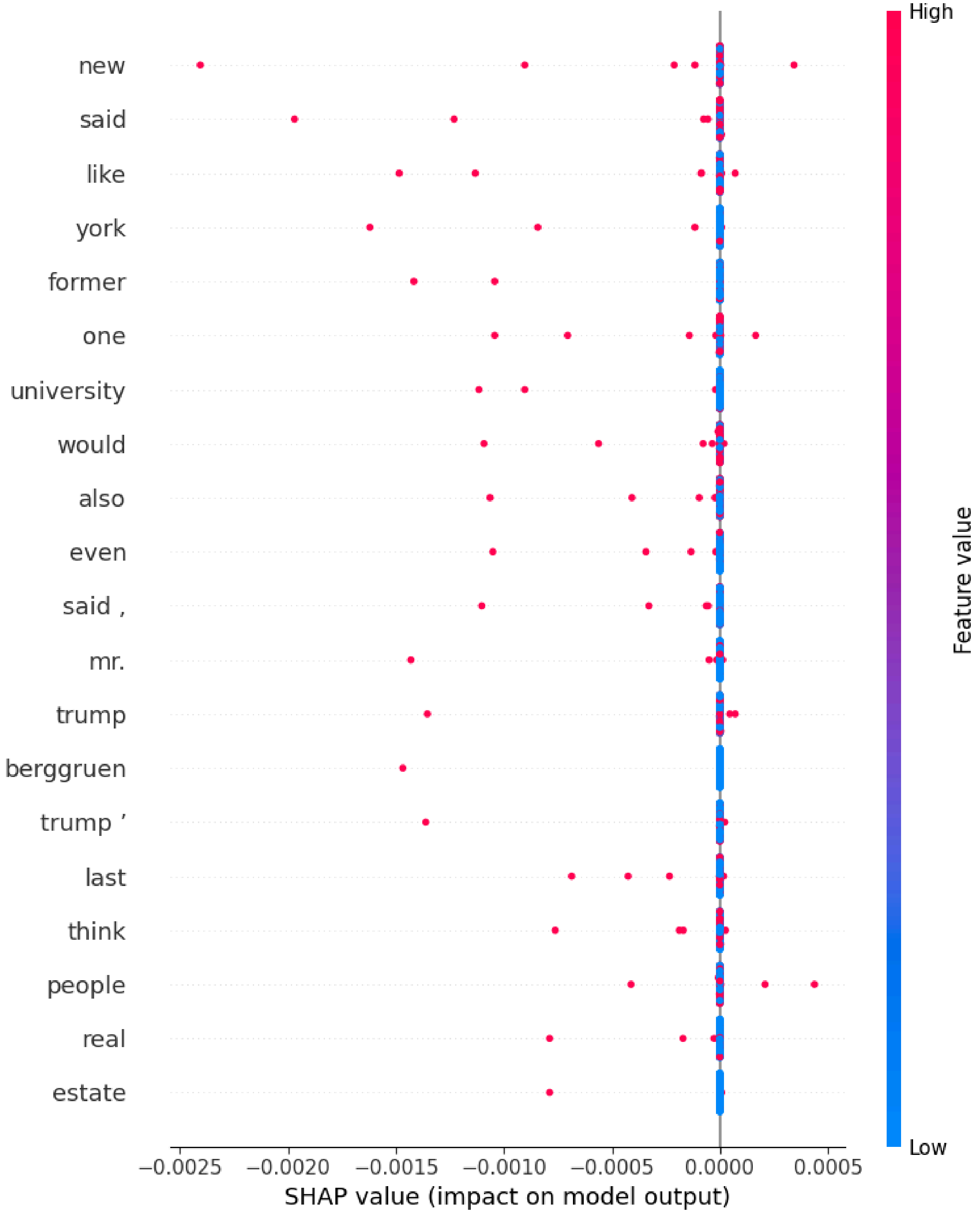

Figure: Global Feature Importance and Impact Analysis of the CNN-LSTM (GloVe) Model

The summary plot displays the top 20 tokens ranked by their contribution to the model's output. The x-axis shows the SHAP value magnitude, which is notably smaller in range compared to the CNN model. Red points indicate the presence of a token, while blue points indicate its absence. It reveals a different distribution of feature importance compared to the CNN-only architecture. The most influential tokens include “new”, “said”, and “like”. A notable characteristic of this plot is the scale of the SHAP values (ranging roughly from -0.0025 to 0.0005), which suggests that individual words in this architecture have a much smaller independent impact on the final prediction, likely due to the LSTM layer processing sequences rather than isolated keywords.

In terms of directionality, there is a strong trend toward negative SHAP values for many top features. When tokens such as “said”, “york”, “former”, and “university” are present (indicated by red points), they consistently drive the prediction to the left (toward Class 0). This contrasts with the previous model where "said" was a positive predictor.

The top feature, “new”, exhibits high variance with mixed polarity; its presence pushes predictions strongly in both negative and positive directions depending on the specific sample. Meanwhile, the token “people” stands out as one of the few top features where presence (red points) tends to contribute positively to the model output (pushing toward Class 1).

5.4.7 SHAP Analysis for CNN-PCA (GloVe)

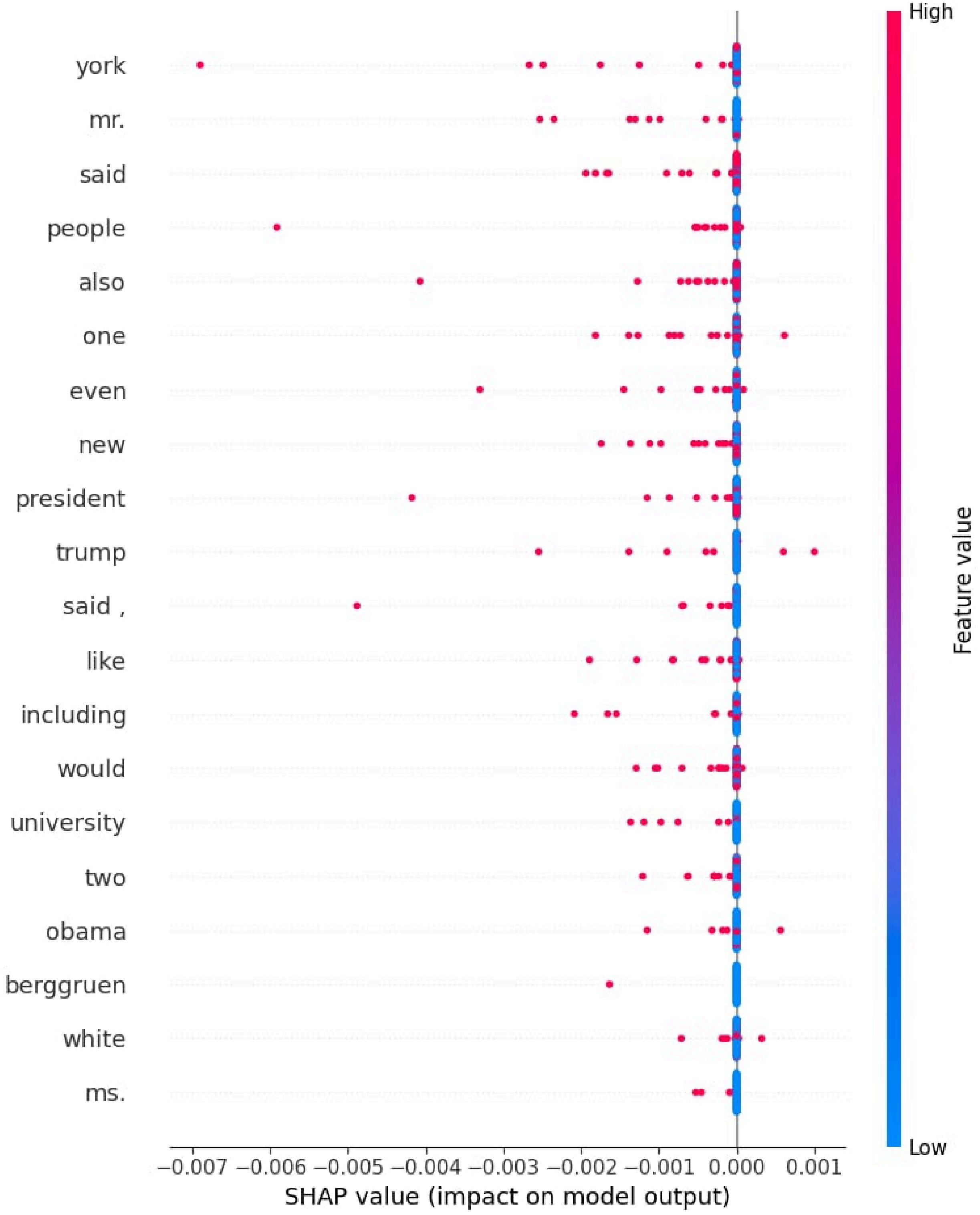

Figure: Global Feature Importance and Impact Analysis of the CNN Model with PCA-Reduced Embeddings

The summary plot ranks the top 20 tokens by their influence on model predictions. The x-axis represents the SHAP value magnitude, which is significantly compressed compared to the full-embedding models. Red points indicate the presence of a token, while blue points indicate its absence. It highlights “york”, “mr.”, and “said” as the most dominant features. Unlike the full-dimension CNN model (Model 3) where “said” was a strong positive predictor, here the presence of “said” (indicated by red points) contributes to negative SHAP values, driving the prediction toward the negative class. Similarly, “york” (likely from “New York”) and “mr.” appear as strong negative indicators when present in the text.

A defining characteristic of this plot is the reduced magnitude of the SHAP values, which mostly fall between -0.007 and 0.001. This suggests that applying PCA to the word embeddings has dampened the discriminative power of individual lexical tokens, forcing the model to rely on smaller, distributed contributions rather than strong keywords.

Most of the top features, including “people”, “also”, and “even”, follow a consistent pattern where their presence exerts a negative influence on the model output. There is a notable lack of strong positive discriminators among the top 20 features, indicating that the model with PCA-reduced inputs primarily learns to identify features associated with the negative class rather than distinct markers for the positive class.

6. AI Safety Analysis

Fake news detection systems present a complex challenge blending technical, ethical, and societal dimensions central to AI safety. At the core, these systems must navigate the tension between maximizing predictive accuracy and ensuring robustness, interpretability, fairness, and trustworthiness under adversarial and evolving real-world conditions.

6.1 Robustness and Generalization

Modern misinformation is increasingly sophisticated, especially with the advent of large language models (LLMs) capable of generating realistic synthetic text, images, and video. Detection models trained on static datasets often fail to generalize against such distributional shifts and LLM-crafted adversarial inputs, leading to false negatives and system failures

6.2 Interpretability and Transparency

The black-box nature of many state-of-the-art models impairs trust and accountability. Explainability techniques such as feature attribution, attention visualization, and mechanistic interpretability can provide actionable insights to human moderators, facilitating corrective interventions when models err or behave unexpectedly (Bereska & Gavves, 2024). Transparent decision processes also support ethical auditability(Trilateral Research,2025), enabling organizations to identify and address biases or vulnerabilities stemming from training data or algorithmic choices (IBM, 2024).

6.3 Fairness and Bias Mitigation

Biases in datasets and models can amplify harmful stereotypes or disproportionately suppress accurate information from marginalized groups, undermining social equity and user trust (Barman et al., 2024; EDPS, 2018). Incorporating fairness-aware training objectives, diverse and representative datasets, and continuous bias auditing is essential for equitable misinformation detection. This is especially critical given that AI models have exhibited differential robustness across demographic and ideological lines (Berman et al., 2015).

6.4 Adversarial and Manipulation Risks

Sophisticated actors exploit vulnerabilities to evade detection or even weaponize AI to generate mass disinformation campaigns. Model jailbreaking, prompt engineering, and coordination across platforms complicate mitigation (The Conversation, 2025). Multi-layered safety architectures with ongoing red teaming, monitoring, and rapid response protocols are necessary to address dynamic threat landscapes (ACIG Journal, 2024; GlobalSign, 2025).

6.5 Ethical Governance and Accountability

AI safety goes beyond ensuring technical robustness; it also requires governance mechanisms that promote responsible development and deployment. Key elements include thorough documentation of dataset origins, transparent reporting on performance, maintaining audit trails, providing avenues for user recourse, and adhering to legal and ethical standards. Additionally, protecting sensitive user data and implementing secure model deployment frameworks are vital to prevent misuse and exploitation. These practices collectively ensure accountability and trustworthiness in AI systems. Privacy safeguards around sensitive user data and secure model deployment frameworks are also critical to prevent misuse or exploitation (IBM, 2024).

6.6 Balancing Accuracy and Safety

Achieving high detection accuracy alone is insufficient. Overconfidence in flawed predictions risks amplifying misinformation or silencing legitimate discourse (Qazi et al., 2025). Systems must balance predictive performance with calibrated uncertainty, explainability, and fail-safe behaviors that enable human oversight and intervention in ambiguous or high-stakes scenarios (Bereska & Gavves, 2024).

7. Limitations

The quality and construction of datasets constitute a core challenge in fake news detection research, as these datasets fundamentally shape model performance and generalization capabilities. Common approaches for creating fake news datasets typically fall into three categories: expert-driven fact-checking, crowd-sourced annotations, and computational or automatic collection based on heuristic or model-driven frameworks.

Expert-driven datasets rely on human domain experts who rigorously verify claims and assign veracity labels based on detailed investigations. While highly reliable, this method is labor-intensive, costly, and thus inherently limited in scale and timeliness (Hamed et al., 2023). Crowd-sourced datasets expand scale by recruiting a wide user base for claim verification, but they suffer from variable annotation quality, inconsistency, and susceptibility to social or political biases (Hamed et al., 2023). Computational approaches leverage automated pipelines combining keyword filters, source credibility heuristics, or distant supervision but often introduce noise and label uncertainty due to the indirect nature of labeling (Verma et al., 2021).

These established dataset creation practices face significant challenges in the current Large Language Models (LLMs) era. LLMs can generate highly plausible fake news that mimics human writing styles and seamlessly integrates factual and fabricated content, making existing datasets less representative of current misinformation forms (Sallami et al., 2024). As a result, models trained on legacy datasets struggle to detect LLM-generated misinformation, suffering from decreased accuracy and increased false negatives. Additionally, few datasets incorporate multimodal information such as images or videos, which are increasingly common in social media misinformation, leaving a critical modality gap unaddressed (Hamed et al., 2023).

Furthermore, the majority of current fake news datasets present binary or coarse-grained labels that fail to encapsulate nuanced misinformation types, such as partially true, satire, or propaganda. This simplification impedes the training of models that can understand subtle distortions or frame-shifting biases pervasive in LLM-generated content (Sallami et al., 2024).

The rapid evolution of misinformation tactics facilitated by LLMs demands continual dataset updates combined with synthetic data augmentation techniques, such as leveraging LLMs themselves to generate challenging adversarial fake news samples (Tong et al., 2025). However, safeguarding against dataset contamination while ensuring diversity and realism remains an open problem.

Current datasets and models frequently lack fine-grained labels that capture the nuanced spectrum of misinformation, limiting the capacity to calibrate system confidence and abstain when uncertainty is high. This shortfall hampers safe deployment, as unchecked predictions may lead to harmful downstream consequences. Balancing predictive accuracy with reliability, interpretability, and uncertainty estimation is thus critical.

Furthermore, the ethical governance of AI requires transparent dataset provenance, auditability, and ongoing monitoring to detect and mitigate biases and fairness issues inherent in data collection and annotation processes. Without these safeguards, biases in fake news datasets can perpetuate societal harms, disproportionately impacting marginalized communities.

Addressing these limitations demands future research focused on constructing dynamic, multimodal, and adversarially robust datasets that reflect evolving misinformation tactics, including those generated by LLMs. Integrating human-in-the-loop frameworks and expanding interpretability tools will be vital for ensuring trustworthy systems that align with societal values and safety objectives.

In summary, while existing fake news datasets have enabled significant progress, they exhibit limitations in scale, representativeness, modality coverage, and label granularity. These limitations are amplified in the emerging LLM context. Addressing these gaps is critical to developing AI systems that remain safe, trustworthy, and effective against the changing landscape of misinformation.

8. Conclusion

9. References

Alshuwaier, Faisal A., and Fawaz A. Alsulaiman. "Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review and Future Perspectives." *Computers*, vol. 14, no. 9, 16 Sept. 2025, pp. 394–394, www.mdpi.com/3501106, https://doi.org/10.3390/computers14090394. Accessed 29 Nov. 2025.

Bereska, Leonard, and Efstratios Gavves. "Mechanistic Interpretability for AI Safety -- a Review." *TMLR*, 2024, arxiv.org/abs/2404.14082. Accessed 29 Nov. 2025.

Bubeck, Sébastien, et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." *ArXiv*, 24 Mar. 2023, arxiv.org/abs/2303.12712, https://doi.org/10.48550/arXiv.2303.12712. Accessed 29 Nov. 2025.

Carlsmith, Joseph. “Is Power-Seeking AI an Existential Risk?” *ArXiv*, 2021, https://arxiv.org/pdf/2206.13353. Accessed 29 Nov. 2025.

Chen, Tianqi, and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, vol. 1, no. 1, 13 Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

Choucair, Cierra. “IBM Launches Unified Data Security Platform to Safeguard against AI and Quantum Risks IBM Launches Unified Data Security Platform to Safeguard against AI and Quantum Risks.” *The Quantum Insider*, 22 Oct. 2024, thequantuminsider.com/2024/10/22/ibm-launches-unified-data-security-platform-to-safeguard-against-ai-and-quantum-risks/. Accessed 28 Nov. 2025.

Dataintelo, and Raksha Sharma. “AI Red Teaming Market Research Report 2033.” *Dataintelo.com*, 30 Sept. 2025, dataintelo.com/report/ai-red-teaming-market. Accessed 28 Nov. 2025.

Fitzgerald, Laura. “AI in the Role of Combating Misinformation | Pindrop.” *Pindrop*, 13 May 2025, www.pindrop.com/article/ai-in-the-role-of-combating-misinformation/.

Hamed, Suhaib Kh. , et al. “A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated with the Dataset, Feature Representation, and Data Fusion.” *Heliyon*, vol. 9, no. 10, 1 Oct. 2023, pp. e20382–e20382, https://doi.org/10.1016/j.heliyon.2023.e20382.

Hendrycks, Dan, et al. “Aligning AI with Shared Human Values.” *ICLR 2021*, 2021, https://arxiv.org/pdf/2008.02275. Accessed 29 Nov. 2025.

IBM. “Explainable AI.” *IBM*, 29 Mar. 2023, www.ibm.com/think/topics/explainable-ai.

Lundberg, Scott M., et al. “Explainable AI for Trees: From Local Explanations to Global Understanding.” *ArXiv*, 11 May 2019, arxiv.org/abs/1905.04610. Accessed 29 Nov. 2025.

Lundberg, Scott, and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” *ArXiv*, 24 Nov. 2017, arxiv.org/abs/1705.07874.

Machová, Kristína , et al. “Analysis of the Effect of Attention Mechanism on the Accuracy of Deep Learning Models for Fake News Detection.” *Big Data and Cognitive Computing*, vol. 9, no. 9, 4 Sept. 2025, pp. 230–230, www.mdpi.com/2504-2289/9/9/230, https://doi.org/10.3390/bdcc9090230. Accessed 30 Nov. 2025.

Molnar, Christoph. *Interpretable Machine Learning a Guide for Making Black Box Models Explainable*. 2019.

Mouratidis, Despoina, et al. “From Misinformation to Insight: Machine Learning Strategies for Fake News Detection.” *Information*, vol. 16, no. 3, 28 Feb. 2025, p. 189, www.mdpi.com/2078-2489/16/3/189, https://doi.org/10.3390/info16030189.

Nawrocki, Mateusz, and Joanna Kołodziej. “Vulnerabilities of Web Applications: Good Practices and New Trends.” *Applied Cybersecurity & Internet Governance*, vol. 3, no. 2, 25 Dec. 2024, www.acigjournal.com/Vulnerabilities-of-Web-Applications-Good-Practices-and-New-Trends,199521,0,2.html#references, https://doi.org/10.60097/acig/199521. Accessed 29 Nov. 2025.

Nele Põldvere, et al. “The PolitiFact-Oslo Corpus: A New Dataset for Fake News Analysis and Detection.” *Information*, vol. 14, no. 12, 23 Nov. 2023, pp. 627–627, https://doi.org/10.3390/info14120627. Accessed 29 Jan. 2024.

Perez, Ethan, et al. “Red Teaming Language Models with Language Models.” *ArXiv*, 7 Feb. 2022, arxiv.org/abs/2202.03286, https://doi.org/10.48550/arXiv.2202.03286. Accessed 29 Nov. 2025.

Qazi, Ihsan A, et al. “Scaling Truth: The Confidence Paradox in AI Fact-Checking.” *ArXiv*, 2025, arxiv.org/abs/2509.08803. Accessed 28 Nov. 2025.

Ribeiro, Marco Tulio, et al. ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier.” *ArXiv.org*, 16 Feb. 2016, arxiv.org/abs/1602.04938.

Rogers, Anna, et al. “A Primer in BERTology: What We Know about How BERT Works.” *ArXiv*, 9 Nov. 2020, arxiv.org/abs/2002.12327, https://doi.org/10.48550/arXiv.2002.12327.

Rudin, Cynthia. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature Machine Intelligence*, vol. 1, 21 Sept. 2019, pp. 206–215, arxiv.org/abs/1811.10154.

Sahan Perera-Merry. “Trust as the Real Currency of AI.” *Trilateral Research*, 25 Sept. 2025, trilateralresearch.com/responsible-ai/trust-as-the-real-currency-of-ai. Accessed 28 Nov. 2025.

Sallami, Dorsaf, et al. "From Deception to Detection: The Dual Roles of Large Language Models in Fake News." *ArXiv*, 2024, arxiv.org/abs/2409.17416.

Shu, Kai, et al. "Fake News Detection on Social Media: A Data Mining Perspective." *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, 1 Sept. 2017, pp. 22–36, dl.acm.org/doi/abs/10.1145/3137597.3137600, https://doi.org/10.1145/3137597.3137600.

---. "Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements." *ArXiv*, 2020, arxiv.org/abs/2001.00623. Accessed 20 Sept. 2024.

Tian, Yexin, et al. "An Empirical Comparison of Machine Learning and Deep Learning Models for Automated Fake News Detection." *Mathematics*, vol. 13, no. 13, 25 June 2025, p. 2086, www.mdpi.com/2227-7390/13/13/2086, https://doi.org/10.3390/math13132086.

Tong, Zhao, et al. "Generate First, Then Sample: Enhancing Fake News Detection with LLM-Augmented Reinforced Sampling." *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers, no. 979-8-89176-251-0, 2025, pp. 24276–24290, aclanthology.org/2025.acl-long.1182/, https://doi.org/10.18653/v1/2025.acl-long.1182. Accessed 28 Nov. 2025.

Verma, Pawan Kumar, et al. "WELFake: Word Embedding over Linguistic Features for Fake News Detection." *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, Aug. 2021, pp. 881–893, https://doi.org/10.1109/tcss.2021.3068519.

Xu, Xiaochuan, et al. "A Hybrid Attention Framework for Fake News Detection with Large Language Models." *ArXiv.org*, 2025, arxiv.org/abs/2501.11967.

Zhou, Xinyi, and Reza Zafarani. "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities." *ACM Computing Surveys*, vol. 53, no. 5, 7 May 2020, https://doi.org/10.1145/3395046.

Lundberg, Scott, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *ArXiv*, 24 Nov. 2017, arxiv.org/abs/1705.07874.

Molnar, Christoph. *Interpretable Machine Learning a Guide for Making Black Box Models Explainable*. 2019.

Mouratidis, Despoina, et al. "From Misinformation to Insight: Machine Learning Strategies for Fake News Detection." *Information*, vol. 16, no. 3, 28 Feb. 2025, p. 189, www.mdpi.com/2078-2489/16/3/189, https://doi.org/10.3390/info16030189.

Nawrocki, Mateusz, and Joanna Kołodziej. “Vulnerabilities of Web Applications: Good Practices and New Trends.” *Applied Cybersecurity & Internet Governance*, vol. 3, no. 2, 25 Dec. 2024, www.acigjournal.com/Vulnerabilities-of-Web-Applications-Good-Practices-and-New-Trends,199521,0,2.html#references, https://doi.org/10.60097/acig/199521. Accessed 29 Nov. 2025.

Nele Põldvere, et al. “The PolitiFact-Oslo Corpus: A New Dataset for Fake News Analysis and Detection.” *Information*, vol. 14, no. 12, 23 Nov. 2023, pp. 627–627, https://doi.org/10.3390/info14120627. Accessed 29 Jan. 2024.

Perez, Ethan, et al. “Red Teaming Language Models with Language Models.” *ArXiv*, 7 Feb. 2022, arxiv.org/abs/2202.03286, https://doi.org/10.48550/arXiv.2202.03286. Accessed 29 Nov. 2025.

Qazi, Ihsan A, et al. “Scaling Truth: The Confidence Paradox in AI Fact-Checking.” *ArXiv*, 2025, arxiv.org/abs/2509.08803. Accessed 28 Nov. 2025.

Ribeiro, Marco Tulio, et al. ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier.” *ArXiv.org*, 16 Feb. 2016, arxiv.org/abs/1602.04938.

Rogers, Anna, et al. “A Primer in BERTology: What We Know about How BERT Works.” *ArXiv*, 9 Nov. 2020, arxiv.org/abs/2002.12327, https://doi.org/10.48550/arXiv.2002.12327.

Rudin, Cynthia. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature Machine Intelligence*, vol. 1, 21 Sept. 2019, pp. 206–215, arxiv.org/abs/1811.10154.

Sahan Perera-Merry. “Trust as the Real Currency of AI.” *Trilateral Research*, 25 Sept. 2025, trilateralresearch.com/responsible-ai/trust-as-the-real-currency-of-ai. Accessed 28 Nov. 2025.

Sallami, Dorsaf, et al. “From Deception to Detection: The Dual Roles of Large Language Models in Fake News.” *ArXiv*, 2024, arxiv.org/abs/2409.17416.

Shu, Kai, et al. “Fake News Detection on Social Media: A Data Mining Perspective.” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, 1 Sept. 2017, pp. 22–36, dl.acm.org/doi/abs/10.1145/3137597.3137600, https://doi.org/10.1145/3137597.3137600.

---. “Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements.” *ArXiv*, 2020, arxiv.org/abs/2001.00623. Accessed 20 Sept. 2024.

Tian, Yexin, et al. “An Empirical Comparison of Machine Learning and Deep Learning Models for Automated Fake News Detection.” *Mathematics*, vol. 13, no. 13, 25 June 2025, p. 2086, www.mdpi.com/2227-7390/13/13/2086, https://doi.org/10.3390/math13132086.

Tong, Zhao, et al. “Generate First, Then Sample: Enhancing Fake News Detection with LLM-Augmented Reinforced Sampling.” *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers, no. 979-8-89176-251-0, 2025, pp. 24276–24290, aclanthology.org/2025.acl-long.1182/, https://doi.org/10.18653/v1/2025.acl-long.1182. Accessed 28 Nov. 2025.

Verma, Pawan Kumar, et al. “WELFake: Word Embedding over Linguistic Features for Fake News Detection.” *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, Aug. 2021, pp. 881–893, https://doi.org/10.1109/tcss.2021.3068519.

Xu, Xiaochuan, et al. “A Hybrid Attention Framework for Fake News Detection with Large Language Models.” *ArXiv.org*, 2025, arxiv.org/abs/2501.11967.

Zhou, Xinyi, and Reza Zafarani. “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities.” *ACM Computing Surveys*, vol. 53, no. 5, 7 May 2020, https://doi.org/10.1145/3395046.