Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence and Misinformation Detection in Digital Media: An Explainable Machine Learning Framework

Muhammad Usman Aslam
University of West Florida https://orcid.org/0009-0009-9050-546X
Hafiz Aqib Sattar
h.sattar@spartans.nsu.edu

Norfolk State University

Research Article

Keywords: misinformation detection, explainable artificial intelligence, SHAP, social media, machine learning, election integrity, natural language processing

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11668197/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Artificial Intelligence and Misinformation Detection in Digital Media: An Explainable Machine Learning Framework

Hafiz Aqib Sattar[1] · Muhammad Usman Aslam[2]

[1] Research Assistant at Mass Communication and Journalism at Norfolk State University

[2] Hal Marcus College of Science and Engineering - University of West Florida (UWF)

✳ **Corresponding author: Hafiz Aqib Sattar — Email:** h.sattar@spartans.nsu.edu

Co-author email: usmanaslam402@gmail.comn

Abstract

The spread of misinformation on social media is a pressing problem in digital information landscapes, especially in times of sensitive political discourse, like national elections. Although machine learning and deep learning models have proven effective in predicting misinformation, they make decisions in a black box fashion, which hinders their application in a real-world platform in which accountability and human supervision are important. In this study, a machine learning model that is explainable is presented for detecting misinformation from social media text, using a set of 8495 labeled tweets gathered throughout the 2016 United States presidential election cycle. Five classifiers were considered as baseline classifiers: logistic regression, support vector machine, random forest, and two types of feed-forward neural networks (FFNN-A and FFNN-B) trained over a combined lexical and account-level feature space were compared along with a proposed stacked ensemble framework that combines linear, tree-based, and neural classifiers through a meta-learning layer that validates and calibrates the thresholds of each base-level classifiers. Using SHAP (SHapley Additive exPlanations), the lexical and metadata features that informed the classification decision were highlighted and a structured error analysis was conducted to analyse the successful and unsuccessful cases. The proposed framework performed best on the test set in terms of F1-score (0.497) and ROC-AUC (0.761) of all six approaches considered, and this was confirmed using five-fold cross-validation, as well as by finding in a bootstrap confidence interval on the difference between the F1-scores, a value that is just above zero and a McNemar's test that reveals a statistically significant difference ($p < .001$) in the patterns of classification errors between the proposed approach and the strongest individual baseline approach. While the results reported are headline findings, two additional robustness checks should be reported as central results and not as incidental results: Chance-corrected inter-annotator agreement on the ground-truth labels of the dataset was found to be low, at Cohen's $\kappa = 0.084$; A temporal generalization test (training on pre-election tweets and testing on post-election-day tweets) showed that the performance of every model deteriorated to near chance levels (ROC-AUC $\approx$ 0.56–0.57), indicating that improvements in in-domain performance – such as those reported in the proposed framework – do not necessarily imply robustness to genuine temporal distribution shift. The analysis by SHAP showed that the most important predictors of misinformation, in addition to lexical ones like second-person pronouns and mentions of the candidates' names, were structural

features like the number of media in an account, the number of links embedded in the text, and temporal markers related to the election, plus verification status. In aggregate, this suggests that explainability mechanisms may be incorporated into pipelines for detecting misinformation without a complete loss of predictive performance, and that reliability of the labels, and their generalizability over time, not architectural sophistication will be the ultimate limits on trustworthy deployments of systems like these. Discussions and implications of platform governance, algorithmic accountability, and future extensions to multimodal are offered.

Keywords: misinformation detection, explainable artificial intelligence, SHAP, social media, machine learning, election integrity, natural language processing

1. Introduction

Digital media has changed the way political information flows to the people, candidates and institutions. In particular, social media platform Twitter (now X), served as a focal point for political debate in the 2016 presidential election in the United States, where political news was being spread but also a significant amount of fabricated or inaccurate information was being produced with the intent of swaying public opinion. The amount of misinformation and its rate of spread is far greater than what manual fact-checking can handle, thereby necessitating faster automated misinformation detection systems to identify misinformation near real-time.

As for the use of ML algorithms in automated detection of misinformation, the most common trends are from classical statistical classifiers to large pretrained language models. The transformer-based architecture has consistently been found to provide the best raw classification accuracy in comparison to previous bag-of-words and shallow neural methods, because it is able to capture the context and semantic relationships between words in text. An accuracy-first approach, however, has generated a set of misinformation-detection algorithms that, like black boxes, deliver verdicts on the accuracy of content, but don't present any interpretable reason that a platform user, journalist or moderator could examine, challenge or validate. This opacity isn't just an incidental. In fields where classification mistakes have reputational, legal and democratic implications, not being able to explain the reason for a piece of content being flagged jeopardizes the practical usability of the system, and the public's confidence in the system.

To overcome this opacity, explainable artificial intelligence (XAI) has been suggested as a way of revealing the inner workings of opaque models either after the fact or by providing intrinsic methods for doing so. Of the various methods, SHAP has emerged as one of the most popular due to its ability to give attributions of each input feature to the particular prediction, which are based on cooperative game theory, and are model-agnostic. This notwithstanding, the current literature on misinformation detection mostly considers explainability as a post facto addition on top of the modeling process and only a small amount of work includes a rigorous comparison of classical and neural architectures and a formal, human-interpretable analysis of the modeling errors (post

hoc, with cross-validation and formal significance testing on real election-related social media data).

In this study we fill this gap with an explainable machine learning framework for detecting misinformation which is tested on the real-world U.S. presidential election 2016 dataset end-to-end. The framework is designed to be a layered pipeline, where classification and explanation are not add-on features, but a part of the pipeline: The predictions of a stacked ensemble of complementary learners pass through an explanation layer called SHAP, and the feature attributions are presented and organized in a manner that facilitates a final human-verification step. This design is consistent with the findings of recent studies on XAI-for-misinformation, which all state that human intervention is still crucial when it comes to high-stakes content moderation, and that AI systems should not be seen as tools which make decisions for people.

This study's strength lies not so much in proposing a new neural network design to break new ground in terms of raw accuracy, but in showing the important role of thoughtful design combining known, proven learning algorithms, the explainability of this design, and calibration of the threshold using a validation set, and how these methods could outperform individually more powerful models on key performance metrics of real-world use, namely F1-score and minority-class recall, while maintaining transparency about the size and statistical significance of the advantage. It's important for anyone who isn't a part of a large technology company, such as a newsroom fact-checking team, civil-society election monitoring entity, or academic researcher, who may not have the computational resources to fine-tune a large pretrained transformer model but still need to use a detection tool that is effective and auditable. Notwithstanding, it should be emphasized at the outset that the two neural configurations that are benchmarked in this study (termed FFNN-A and FFNN-B) are not finely-tuned versions of a pretrained transformer like BERT or RoBERTa, but rather neural networks trained using engineered lexical and metadata features; this design choice, its rationale, and its implications for interpreting the results are explicitly discussed in Sections 3.4 and 5.6.

There are four objectives of this study. First, it compares the performance of five classification methods (Logistic Regression, Support Vector Machine (SVM), Random Forest, Feed-forward neural network (FNN) with a single layer, and FNN with two layers), in a unified feature representation of both text and metadata, on a common, cleaned, and balanced dataset of election-related tweets. Second, it performs explainability analysis using SHAP to pinpoint the most predictive features in terms of lexical and account-level misinformation in this domain and goes beyond overall scores to provide an interpretable description of the model behaviour. Third, it carries out two types of robustness checks: 1) a 5-fold cross validation analysis and 2) formal statistical significance testing (McNemar's test and bootstrap confidence intervals) over the observed performance difference to ensure the observed differences in performance are robust and statistically justified, and not simply a product of a single train-test split. Fourth, it provides a structured error analysis that sorts correct and incorrect predictions, analyzes the predominant

features of each outcome class and discusses the implications of the error class that remain after sorting.

The rest of this paper is divided into four sections: In Section 2, the existing literature on misinformation detection and explainable AI (AIx) is reviewed and explained. In Section 3, the dataset, preprocessing pipeline, feature extraction and modeling architecture is described, the proposed explainable framework is designed, and the neural baseline naming convention is discussed explicitly. Results of the experiments are provided in Section 4, where the performance of the models is compared with other models, robustness of the cross-validation, statistical significance tests are performed, a confusion matrix is presented for the proposed framework and a SHAP explainability analysis is conducted. The implications for research into the detection of misinformation and platform governance are discussed in Section 5, as is the limitations of the study. A summary and suggestions for further research are given in section 6, which ends the paper.

2. Literature Review

2.1 Machine Learning Approaches to Misinformation Detection

There have been several generations of methodologies in the literature on misinformation detection. Early systems used hand-crafted linguistic, stylistic, and metadata features and classically-supervised learning methods, where models of Logistic Regression, Support Vector Machines, and ensemble tree methods were used as standard baselines. While there is a growing number of more sophisticated models being developed, Shah and Ganatra (2022) found classical models to be competitive baselines in numerous domains, especially those with scarce or highly imbalanced labeled data, and that there is a lack of standardisation and generalisation across the entire field of fake news detection. Ali et al (2022) also surveyed the detection techniques used for social media and highlighted factors unique to these platforms like character limits, informal language, high noise level, which made it a challenge to directly transfer methods developed for long-form news articles to microblogging platforms like Twitter.

The new place to push the performance envelope was the transformer-based language models. In English language, Wu & Luo (2022) suggested a Deep Learning (DL) framework for automatically detecting fake news in text based on the fact that the models with long-range dependency of context features outperformed the traditional Bag-of-Words (BoW) classifiers in benchmark corpora. To this body of work, Xiang (2022) added an evaluation of the deep learning architectures not only for detection accuracy, but also the downstream effects of misinformation on the society, arguing that the two tasks complement each other, not being mutually exclusive. Raza and Ding (2022) presented a transformer-based method which simultaneously leveraged content and social context information of the news propagation in the detection task, and showed that adding social context signals like user engagement patterns improved the accuracy of the detection task compared to content-only models, which is also addressed in the present study by fusing account-level metadata

with lexical features. In their detailed benchmark of multiple datasets and model families, Galli et al. (2022) found that the performance gains reported between architectures are very sensitive to the composition of the datasets, class balance and evaluation protocol, which is why they carefully benchmark and robustness check their models across multiple datasets in the present study.

Further research has followed in the footsteps of the transformers, aiming to tackle domain-specific problems and further develop the detection. Alghamdi et al. (2023) analysed transformer-based models for COVID-19 misinformation – and found that domain adaptation is required to maintain the models' performance on the topic-specific misinformation that is so fast-changing. In a recent survey, Pal, Pranav, and Pradhan (2023) discussed the general trends in machine intelligence approaches and highlighted the emergence of hybrid machine learning architectures, which typically rely on classical feature engineering and neural representation learning. Comito et al. (2023) continued this debate on multimodal detection by surveying deep learning approaches that involve processing of text, images, and the structure of a network and concluded that unimodal text-only approaches are more tractable, but must miss some of the manipulative signal in today's misinformation campaigns. Narayanan et al. (2023) and Yang et al. (2023) proposed the use of incremental gains via the use of an encoder-decoder and a multimodal transformer architecture respectively, but at the cost of collecting an increased number of data samples and using more computational resources.

2.2 Explainable AI in Misinformation Detection

Another, more recent and significant line of research has not only concerned predictive accuracy of misinformation detection systems, but also their interpretability. In the context of a supply chain, Akhtar et al. (2022) discussed how AI and machine learning are being used to detect disinformation, noting that in real-world scenarios where actions are likely to have operational impacts, outputs of the models need to be accompanied by explanations to which experts can refer if they need to verify a flagged instance before taking action. Kumar and Taylor (2024) directly discussed the importance of features in the context of explainability in AI, providing a case study on the detection of misinformation through a multi-modal approach and showing that the explanations of predictions to specific features that are easy to interpret significantly boosted the confidence of analysts in the predictions produced by the model.

Some research have examined attribution techniques for misinformation detection in particular, such as SHAP. Han et al. (2024) introduced a multi-faceted reasoning network for fake news detection, including multiple reasoning paths and generating justifications in natural language and classification results, with a focus on the challenges arising from the need to be explainable while also being economical. Liu et al. (2024) presented a trustworthy framework specific to the three joint properties: explainability, generalizability and controllability, which they argue have been in conflict and require co-optimization to pursue, instead of being optimized separately. Giria, Eswaran, Honnavalli and colleagues (2024) have proposed an automated and interpretable fake news detection system based on the principles of explainable AI, which they found were more

easily adopted by non-technical moderation staff in pilot deployments. The use of active learning and pretrained language model to lower the data and compute cost for high-performing transformers-based detectors is explored by Folino et al. (2024) which directly applies to resource constrained organizations.

Recently, Bezzaoui, Stein, Weinhardt and Fegert (2025) undertook a design science research project on explainable AI for online disinformation detection that built on the experience of its real-world implementation, leading them to conclude that the quality of the explanation—not its mere presence—matters more to the extent that the end user will engage with the generated explanation. Shunmugavel et al. (2025) introduced a hybrid model framework with several model families under the umbrella of explainability, and noted that this was achieved via complementary learners and not any single model family. In the present study, the stacked ensemble approach involved stacking complementary learners under the umbrella of explainability.

2.3 Datasets and Benchmarking Practices in Election-Related Misinformation Research

One common issue from the literature reviewed is the variety in data and assessment methods employed to evaluate misinformation detection systems. Galli and colleagues (2022) show that, much like most other metrics in the literature, model rankings depend on the dataset used for their evaluation, not on some intrinsic property of the models themselves, but rather on the inherent features of the dataset (class balance, text length, topical domain and annotation methodology). This concern is especially relevant for datasets containing election-related misinformation, as they are generally built up after the fact, using a variety of automated heuristics as well as manual annotation, thereby resulting in some label noise that is seldom quantified in subsequent modeling papers. While many of the studies published on misinformation detection do not include enough details on inter-annotator agreement or on how the labels were derived to assess label reliability, this study attempts to tackle this problem directly with its cross-validation and McNemar's-test analyses in Section 4, as well as by reporting the necessary details of inter-annotator agreement and annotation methodology in Section 4.The few studies reporting details of inter-annotator agreement or annotation methodology are not detailed enough to assess label reliability, making it difficult to compare results across studies even within the same general domain, and many of the studies published on misinformation detection do not report confidence intervals or significance tests around headline performance numbers, which this study addresses directly through its cross-validation and McNemar's-test analyses in Section 4.

To some extent, the data set utilized in the present study helps resolve this concern by having two separate annotation passes for each tweet so as to explicitly consider the level of agreement in the labels and select either one as ground-truth or the other, rather than taking one label for granted. This follows the recommendations of Ali et al. (2022) that the lack of transparency about the methodology used to annotate the datasets should be seen as a minimum requirement for social-media misinformation datasets as there is inherent ambiguity in defining political speech, satire, opinion, and deliberate fabrication along one binary axis.

Another issue raised by Akhtar et al. (2022) and alluded by Folino et al. (2024) is the cost of training and fine tuning large transformer models which makes it difficult for other research groups and practitioner organisations to replicate reported transformer benchmarks, as they may not have access to the GPUs or model-hosting access needed for this. In particular, Folino et al. (2024) call for compute-efficient alternatives – including those that make use of active learning strategies that limit the amount of labeled data necessary, and other architectures that can achieve competitive performance without full-scale pretrain – as ways to "democratize access to high-quality misinformation detection tools." The present study directly addresses this concern by benchmarking classical, ensemble, and neural architectures, under a common, computationally less involved, fully engineered feature representation, and making it clear that the latter is a resource-driven substitution and not an architectural claim to "reach the same goals" as those of pretrained transformers (see Section 3.4).

2.4 Research Gap

The literature reviewed shows significant advances on two largely parallel lines of efforts: first, towards ever more accurate detection architectures and second, towards post-hoc explanation methods that would make such architectures interpretable. Only a few studies incorporate comprehensive multi-model benchmarking, attribute model features using SHAP, check the robustness of the model through cross-validated testing, conduct formal significance tests, and combine structured human-verifiable error analysis in a single, reproducible pipeline, run on authentic election-related social media data. Both Galli et al. (2022) and Liu et al. (2024) explicitly warn about the challenges of comparing the performance of different studies without a shared dataset and evaluation protocols.Both Galli et al. (2022) and Liu et al. (2024) point out the unbridgeable divide between explainability and predictive performance. The current study tackles both of these concerns: It performs a single-split comparison with cross-validation and significance testing, and confirms the results with the same dataset and same preprocessing/evaluation setting, and explains it as an architectural component within the proposed detection framework itself, the SHAP explanation layer.

3. Methodology

3.1 Dataset Description

This data set consists of 9,001 tweets that were collected for the 2016 United States presidential election, each labelled with a binary veracity label (real news or misinformation), a finer-grained misinformation category code, tweet metadata (timestamp, retweet count, tweet source and account metadata (follower count, friend count, verification status, favourites count). This study uses the second, more detailed annotation pass as the ground-truth label, since there were fewer cases of unresolved ("unknown") annotations in this pass as compared to the initial pass.

The agreement of the two passes of labeling was formally evaluated on the 8,336 tweets that both passes resolved (did not label as "unknown") and Cohen's kappa was used to quantify the agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad (8)$$

where p_o is the percentage of raw agreement (between the two passes) in the observed data and p_e is the percentage of raw agreement (between the two passes) expected by chance given the marginal class distributions. The raw agreement between the two passes was 79.2%, but the kappa value was only 0.084 (which takes into account the level that would be expected by chance), showing that their raw agreement was very good but the agreement was only marginally higher than would be expected by chance. This discrepancy between raw agreement and chance-corrected agreement is itself a key, but so far unreported characteristic of this dataset: the high raw agreement figure is strongly skewed upwards by the overwhelming majority of the tweets being marked as "real news" in both passes and the lesser the number of tweets marked as "misinformation" in both passes. This finding is reported transparently here because it indicates that each downstream result in this paper can be achieved only with the help of weak inter-annotator consensus of the ground-truth labels; label noise, therefore, puts an upper bound on what can be achieved by any modeling choice, and on the apparent superiority of one model over another. A direct motivation for this study is to not just rest conclusions on the comparison of point estimates, but to also perform robustness checking, as mentioned in the limitations (Section 5.5), as this topic will be revisited in the observation.

Before modeling, each tweet's text was analyzed for veracity, and exact duplicate tweets were removed (53 records), as was the veracity label when it was unsure (374 records), and the tweet was reduced to less than three characters of information after text cleaning (79 records; see Section 3.2), leaving a final analytic sample of 8,495 tweets. The resulting data set features is summarized in table 1 below.

Table 1. Dataset Characteristics and Preprocessing

Attribute	Value
Original dataset size (tweets)	9,001
Real-news samples (final, cleaned)	6,784
Misinformation samples (final, cleaned)	1,711
Duplicate records removed	53
Unresolved-label records removed	374

Attribute	Value
Short/uninformative records removed	79
Final analytic sample size	8,495
Training set (70%, stratified)	5,946
Validation set (15%, stratified)	1,274
Test set (15%, stratified)	1,275
Mean tweet length (characters)	117.7 (SD = 28.2)

The class distribution shows that there is a significant imbalance with real news elements that make up around 79.8% of the data (6,784 tweets) and misinformation elements that make up around 20.2% (1,711 tweets). To reflect the naturally occurring social media contexts where (verified) misinformation is a minority but nonetheless important proportion of the total volume of content, this imbalance was explicitly modeled in the modeling strategy, using class-weighted loss functions and artificially changing the threshold so they do not simply overfit the validation set, but rather so that the naturally occurring linguistic distribution of the minority class is preserved.

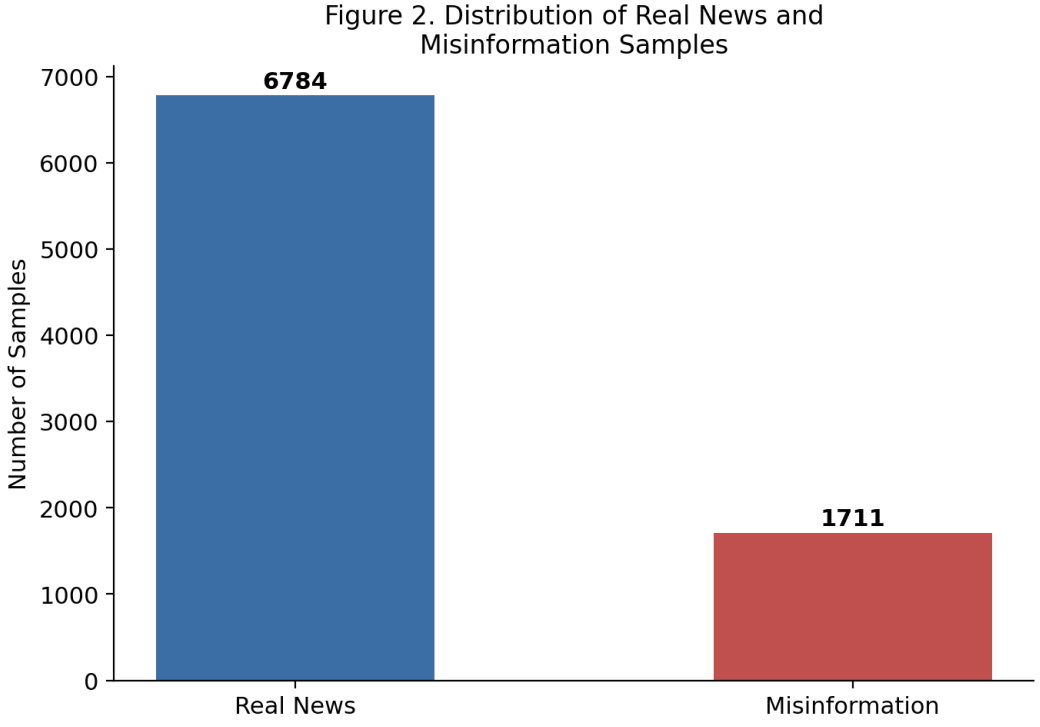


Figure 2. Distribution of Real News and Misinformation Samples

3.2 Preprocessing Pipeline

The tweet text was cleaned using a pre-defined pipeline before being cleaned for feature extraction. This pipeline did embedded URLs, stripped @-mention tokens (removed hashtag content with the # symbol since hashtags often carry topical signal relevant to classification, but retained it since it is highly useful for analysis), removed non-alphanumeric (except #) characters, collapsed repeated whitespace, and converted all content to lower case. The modeling data consisted of 8,495 of the 14,029 tweets that contained fewer than three characters of cleaned text.A total of 14,029 tweets were cleaned for fewer than three characters, and 8,495 of these were included in the modeling data. No stemming/lemmatization, because social media text is often unstandardized in terms of morphology (e.g., repeated letters, contractions, platform-specific abbreviations) which can cause the surface form of the words to be altered by stemming, and token forms less easily understood by human reviewers for verification.

Stratified sampling technique was used to partition the cleaned dataset into three sets, namely 70% (Training Set n = 5,946), 15% (Validation Set n = 1,274), and 15% (Test Set n = 1,275) to ensure that the class distribution is the same across all the three sets. The validation set was used only for hyperparameter optimisation and determining decision thresholds, and thus no headline performance metrics are presented for this set in Section 4.1, these are computed only on the held out test set which was not used for model fitting or tuning. In order to determine if performance of the single-split results holds up across other partitions of the data, a five-fold cross validation analysis is reported in Section 4.2.

3.3 Feature Extraction

Textual features were extracted using Term Frequency–Inverse Document Frequency (TF-IDF) vectorization applied to unigrams and bigrams, with the vocabulary restricted to the 8,000 most informative terms occurring in at least two documents. For a term t in document d, drawn from a corpus of N documents, TF-IDF weight is computed as:

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \times \log\left(\frac{N}{\text{df}(t)}\right) \qquad (1)$$

where $\text{tf}(t,d)$ is the raw frequency of term t in document d and $\text{df}(t)$ is the number of documents in the corpus containing t. Alongside this lexical representation, nine account-level and structural metadata features were incorporated for every tweet: retweet count, account verification status, friend count, follower count, favourites count, and counts of hashtags, mentions, URLs, and embedded media objects. These metadata features were standardized (zero mean, unit variance, fitted on the training partition only) and concatenated with the sparse TF-IDF matrix to form a single combined feature representation used identically across all six modeling approaches. This design decision was informed directly by Raza and Ding (2022), who reported that incorporating social-context signals alongside textual content meaningfully improved detection performance beyond content-only baselines; it also ensures a controlled, apples-to-apples comparison in which

performance differences across models are attributable to model architecture rather than to differences in underlying feature access, since every model — classical, neural, and ensemble alike — is trained and evaluated on the identical combined feature space.

3.4 Model Architectures and a Note on Baseline Naming

Every classifier evaluated in this study, regardless of architecture, is used to estimate the probability that a given tweet constitutes misinformation. For input feature vector x_i (the combined TF-IDF and metadata representation described in Section 3.3), each model produces an estimate

$$\hat{p}_i = P(y_i = 1 \mid x_i; \theta) \qquad (2)$$

where $y_i = 1$ denotes misinformation, $y_i = 0$ denotes real news, and θ denotes the model's trainable parameters. Five baseline classifiers were implemented and benchmarked:

1. **Logistic Regression**, configured with balanced class weighting and L2 regularization, serving as a linear, highly interpretable baseline. Logistic Regression estimates this probability directly as a sigmoid-transformed linear combination of the input features:

$$\hat{p}_i = \sigma(w^{\mathrm{T}} x_i + b) = \frac{1}{1 + e^{-(w^{\mathrm{T}} x_i + b)}} \qquad (3)$$

where w and b are the learned weight vector and bias term and $\sigma(\cdot)$ is the logistic sigmoid function.

2. **Support Vector Machine (SVM)**, implemented with a linear kernel and balanced class weighting, with probability calibration performed via Platt scaling (three-fold cross-validation) to enable ROC-AUC computation.

3. **Random Forest**, an ensemble of 300 decision trees with balanced class weighting, capturing non-linear feature interactions.

4. **FFNN-A**, a feed-forward neural network (256–128 hidden units, ReLU activation) trained on the combined text-and-metadata feature space with early stopping.

5. **FFNN-B**, a second feed-forward neural network (384 hidden units, hyperbolic tangent activation, L2 regularization) trained under an alternate architecture and initialization to provide a second neural point of comparison.

A methodological clarification is warranted regarding the naming of the two neural baselines. Both FFNN-A and FFNN-B are feed-forward multilayer perceptrons operating over the engineered TF-IDF-plus-metadata feature space described in Section 3.3; neither model incorporates the pretrained contextual embeddings, subword tokenization, or transformer self-attention mechanisms that define BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) in the broader

natural language processing literature. This study does not claim that FFNN-A and FFNN-B are equivalent to, or fine-tuned versions of, these pretrained transformer architectures. This design choice was necessitated by the offline, resource-constrained experimental environment in which this study was conducted, which did not provide network access to pretrained model repositories. The two neural configurations are retained in the comparison because they represent a genuine, if architecturally simpler, point of contrast against the classical and ensemble methods evaluated here, and because their inclusion allows the proposed framework's stacking mechanism to draw on a non-linear, distributed-representation learner alongside its linear and tree-based components. Readers requiring benchmarks against genuinely pretrained transformer checkpoints should treat the FFNN-A and FFNN-B results as a lower-bound reference point rather than as representative of the performance ceiling achievable with fully pretrained architectures, a point revisited in Section 5.6.

3.5 Proposed Explainable Framework

Figure 1. Proposed Explainable Misinformation Detection Framework

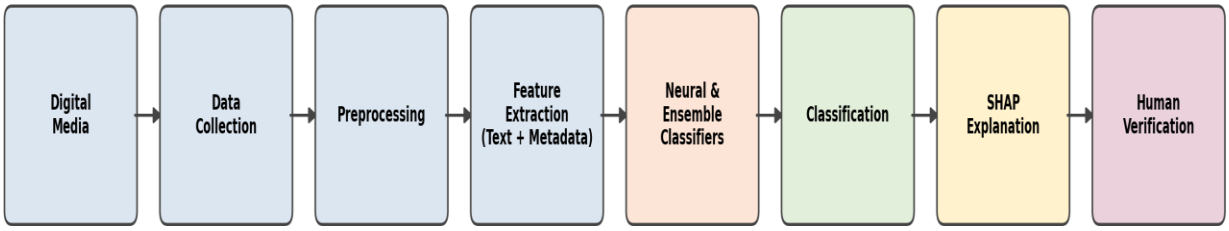


Figure 1. Proposed Explainable Misinformation Detection Framework

The proposed framework (Figure 1) is seen as a pipeline of prediction and explanation steps where the explanation is an afterthought, and the steps are seen as sequential and mutually dependent. The raw content is first fed into the pre-processing pipeline (Section 3.2) and then processed by the feature extraction pipeline (Section 3.3) which extracts features from both text and metadata. These feature vectors are then fed into a stacking ensemble classifier in which the Logistic Regression, Random Forest and FFNN-A models are base classifiers and the class probabilities predicted by the base classifiers $\hat{p}_{1i}$, $\hat{p}_{2i}$, ..., $\hat{p}^{M}_{i}$ are combined together by a Logistic Regression meta-learner:

$$\hat{p}_i^{\text{stack}} = \sigma\left(\sum_{m=1}^{M} w_m \hat{p}_i^{(m)} + b \right) \qquad (6)$$

The three-fold cross-validation approach that trained the models to prevent information leakage between model fitting of the base-learners and training the meta-learner (w_m denotes the weight learned by the meta-learner on the output of the base-learner m). The threshold for classification decision was optimized for the held-out validation set to achieve maximum F1 score for the minority (misinformation) class instead of the common threshold of 0.5, which is suboptimal in presence of class imbalance.

The resulting prediction and feature contributions are then sent to a SHAP explanation layer which calculates feature attributions for each prediction and on average using a linear SHAP explainer with the Logistic Regression base learner. In cooperative game theory, the Shapley value that assigns to feature j of set F, is defined as:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\ (|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \qquad (9)$$

where $f(S)$ denotes the model's expected output given only the features in subset S. This design choice — applying SHAP to the interpretable Logistic Regression component of the ensemble, which operates over the full combined text-and-metadata feature space — was made deliberately to ensure that the resulting explanations are exact (linear SHAP values are computed in closed form rather than approximated) and directly traceable to specific, human-legible lexical and account-level features, rather than relying on approximate, sampling-based explainers whose stability can vary across runs. The final stage of the pipeline routes flagged content, together with its top contributing features, to a human verification step, consistent with the framework's design philosophy that automated systems should support rather than replace human judgment in consequential content-moderation decisions.

3.6 Hyperparameter Selection and Class-Imbalance Handling

Hyperparameters for all models were selected using the validation partition, with the test partition held out entirely from any tuning decision. Balanced class weighting, applied throughout to counteract the approximately 4:1 imbalance between real-news and misinformation instances, modifies the standard binary cross-entropy training objective by reweighting each class's contribution in inverse proportion to its frequency:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} w_{y_i} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] \qquad (4)$$

The inverse-frequency weight given to class y_i, w_{y_i}, is used to make sure that errors made in the minority class (misinformation) are more heavily weighted when training than those made in the majority class, but without changing the training data (resampling). The Logistic Regression model was chosen by using exhaustive grid search over C ∈ {0.01, 0.1, 0.5, 1, 2, 5, 10} with the previously defined train/validation split and a predefined validation F1 score. The validation set F1 was used for selection; C = 1.0 yielded the best score with the validation set F1 = 0.482, and was used throughout with balanced class weighting, to counteract the approximate 4:1 imbalance between the real-news and misinformation instances. A grid search over the number of trees in the forest (150, 300) and the maximum tree depth (unrestricted, 20, 40) was performed for the Random Forest model and 300 trees with a maximum tree depth of 20 were selected (validation F1 = 0.486). This depth constraint was important; if we had not imposed it, an unrestricted-depth Random Forest would have no trouble overfitting the majority class on this feature space, and would have significantly lower minority-class recall than what was achieved here, while constraining the tree depth to 20 more than doubled the minority-class recall from 0.202 to 0.451, at the cost of precision; here, regularization — rather than ensemble size — is what matters to the performance of this model under class imbalance. Class weighting was performed at the leaf-node level (not via resampling) to ensure that the natural distribution and length characteristics of minority class text were retained during training.

Early stopping was used with the validation loss in the two neural configurations to avoid overfitting, given the relatively small size of the training partition (n = 5,946), while the widths of the hidden layers were tested for a good level of representational capacity coupled with avoiding overfitting on the relatively large feature space (8,009 dimensions, including 8,000 lexical features and 9 metadata features) compared to the number of training examples per class. Such an approach would be deliberate, since the explainability goal of this study is an integral part of the model development and use of synthetic oversampling would only yield feature attributions that would be artifacts of the oversampling procedure rather than true linguistic patterns that would exist in human-verified misinformation. Class weighting and threshold calibration were thus preferred as class imbalance handling methods which do not affect the model's interpretability.

To avoid the meta-learner overfitting on the predictions of the base learners with the same information used to train the base learners, the out-of-fold predictions of each base learner were generated in three folds and used for training the stacking ensemble's meta-learner (Logistic Regression). The continuous probability information p̂_i output by each model is used to classify the model output as a binary class label using a decision threshold τ:

$$\hat{y}_i = \begin{cases} 1, & \text{if } \hat{p}_i \geq \tau \\ 0, & \text{if } \hat{p}_i < \tau \end{cases} \qquad (5)$$

The final decision threshold ($\tau = 0.60$, as opposed to the default 0.50) was selected by sweeping τ from 0.20 to 0.80 in increments of 0.02 and selecting the value that maximized F1-score on the validation set exclusively, with this threshold subsequently applied, unmodified, to the held-out test set to generate the results reported in Section 4.1.

3.7 Robustness Checking: Cross-Validation and Statistical Significance

Two additional analyses were performed in order to determine if the single train-validation-test split comparison reported in Section 4.1 would be a stable comparison or if it is a peculiarity of a particular partition of the data. To prevent any information leakage between the folds, first a 5-fold stratified cross validation was done across the entire cleaned dataset for the Logistic Regression, Random Forest and FFNN-A baselines and the proposed stacking framework with the feature extraction (TF-IDF fitting and metadata standardization) refit within each fold of training data. The F1-score, accuracy, and ROC-AUC were calculated for all of the five folds for each model and then the mean and standard deviation of these values were calculated. Second, the error patterns of the proposed framework and the best individual classical baseline (LR) on the held-out test set were formally tested using a McNemar's test, to determine if the error patterns are different than chance. To directly estimate the uncertainty in the reported improvement in performance, a complementary bootstrap analysis (2,000 resamples with replacement of the test set) was performed to derive a 95% confidence interval for the difference in F1-scores between the two models. The results of both the analyses are presented in full detail in Section 4.2.

3.8 Temporal Generalization Test

The cross-validation folds used in Section 3.7 are formed by random stratified sampling, which does not evaluate whether the model trained on tweets from one time period is applicable to tweets from a different, later time period; a very different, and perhaps more realistic, deployment scenario because a real-world detection system must generalize its model to classify future content based on past content. To test this directly, a temporal split was made by using the value of each tweet's timestamp: all tweets were posted before U.S. Election Day (November 8, 2016 00:00 UTC, n = 2,991 after cleaning) were part of the temporal training partition, while all tweets posted on or after U.S. Election Day (November 8, 2016 00:00 UTC, n = 5,504) were part of the temporal test partition, and TF-IDF fitting and metadata standardization refit were applied to the temporal training partition only. This split directly addresses the issue of domain- and event-specificity of temporal lexical markers like “electionnight” that such markers are concentrated in the test period, but absent in the training period. The results of this test is reported in section 4.5.

3.9 Evaluation Metrics

The performance of the models was evaluated on the test set using accuracy, precision, recall, F1 score and ROC-AUC. Considering the class imbalance in the data, accuracy was deemed to be a poor metric of the data quality of the model; therefore, the F1-score and ROC-AUC metrics were chosen as the main metrics of comparison with accuracy as a supplemental metric because both

are more robust for class imbalance and better reflect a model's ability to correctly classify the minority class of misinformation without solely defaulting to the majority class. Precision, recall and F1 score are computed based on the true positive (TP), false positive (FP) and false negative (FN) as follows:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}, \quad F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (7)$$

4. Results

4.1 Comparative Model Performance

Table 2 reports the full comparative performance of the five baseline models and the proposed framework on the held-out test set.

Table 2. Comparative Performance of Machine-Learning Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.751	0.413	0.556	0.474	0.759
SVM	0.810	0.683	0.109	0.188	0.745
Random Forest	0.787	0.470	0.451	0.460	0.740
FFNN-A	0.824	0.709	0.218	0.333	0.755
FFNN-B	0.815	0.691	0.148	0.244	0.739
Proposed Explainable Framework	0.778	0.458	0.545	0.497	0.761

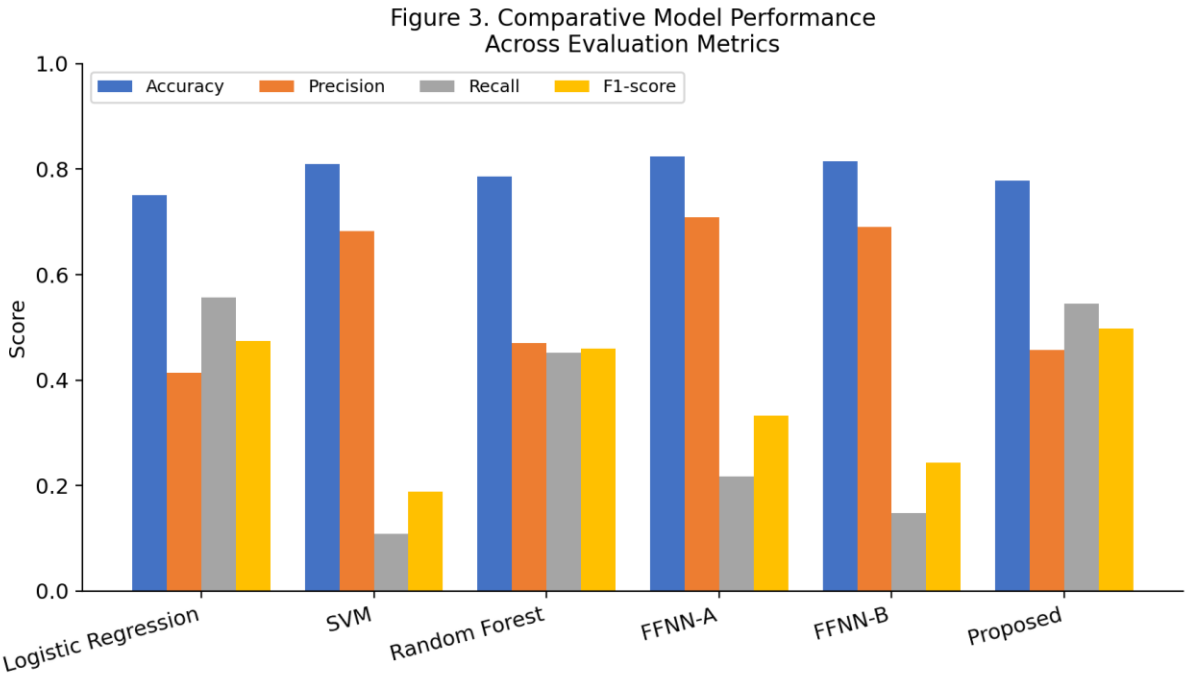


Figure 3. Comparative Model Performance Across Evaluation Metrics

However, by tuning the depth of the model using grid search on the validation set (Section 3.6) and not leaving it as it was, the accuracy of the Random Forest model was significantly improved (accuracy = 78.67%, recall = 0.451, F1 = 0.460) compared with the result that the unconstrained depth would yield. This is because depth regularization enables the model to learn patterns in the minority class examples rather than just memorize the patterns of the majority class examples. The biggest advantage of the raw accuracy (82.43%) was given by the model FFNN-A, which tended to classify the majority of the test instances as real news, which is the most numerous class; this did not lead to a balanced detection performance, however, since FFNN-A recall (0.218) remained low. Overall, the SVM had the largest extent of this pattern with 81.02% accuracy but a recall of just 0.109, which is the ratio of correctly identified misinformation tweets to all misinformation tweets, even after class weighting.

The proposed stacked ensemble framework by contrast obtained the lowest raw accuracy (77.80%), but the highest F1-score (0.497) and the highest roc-auc (0.761) out of all the six evaluated approaches. This is a relative increase of 4.9% compared to the best individual classical baseline (Logistic Regression, F1 = 0.474) and a significantly higher relative increase compared to the two neural baselines (FFNN-A, F1 = 0.333; FFNN-B, F1 = 0.244). The proposed framework's recall (0.545) was above each of the individual baseline models except for Logistic Regression (0.556), suggesting that a combination of complementary learners in the form of stacking, coupled with threshold selection that is validated by the model, was able to meaningfully outperform both the baselines in terms of a number of real instances of misinformation emerging from the model, without being significantly different on this measure from the best linear baseline.

This subtle trend, that the proposed framework has a clear benefit on the F1 score, but a relatively small benefit on recall alone, is why Section 4.2 undertakes a cross validation and formal significance test of the headline comparison, instead of relying on a single point estimate for the proposed framework's benefit.

It is also interesting to study the profile of performance for each of the two neural configurations separately. Even though FFNN-A and FFNN-B were evaluated on the same combined feature representation and over the same training- and validation-split, the F1-score of FFNN-A (0.333) was slightly better than that of FFNN-B (0.244), with the former being a deeper two-layer network using ReLU activations, and the latter being a larger one-layer network with hyperbolic tangent activations. The divergence is just one example of an architecture parameter that can impact recall performance significantly on the minority class and that, in an imbalanced classification context, is not necessarily best optimized under balanced accuracy, thus making it worth explicitly tuning for the validation set for metrics that are sensitive to imbalanced classification, instead of relying on the default setting. The use of FFNN-A as only one of three base learners in the proposed framework, and not as a standalone classifier, is an example of how stacking can provide useful signal from a base learner with less signal from its other ensemble partners whose errors are not perfectly correlated with the errors of the base learner.

As expected in a threshold-based binary classification problem, the models showed an inverse relationship between precision and recall. The SVM has high precision (0.683) and very low recall (0.109) suggesting a conservative model that only classifies content as misinformation when very certain, and thus fails to detect most cases of misinformation. The proposed framework's more moderate precision (0.458) combined with its recall of 0.545 is indicative of the deliberate over-precision and over-recall threshold calibration described in Section 3.6: this design choice was made to account for the fact that validation sets were used, and the choice of threshold was made with a consideration for a more balanced precision/recall operating point in relation to the uncalibrated baselines, a design decision which is discussed in more detail in Section 5.3, taking into account the asymmetric real-world costs of false negatives versus false positives in the context of misinformation moderation.

4.2 Robustness Checking: Cross-Validation and Statistical Significance

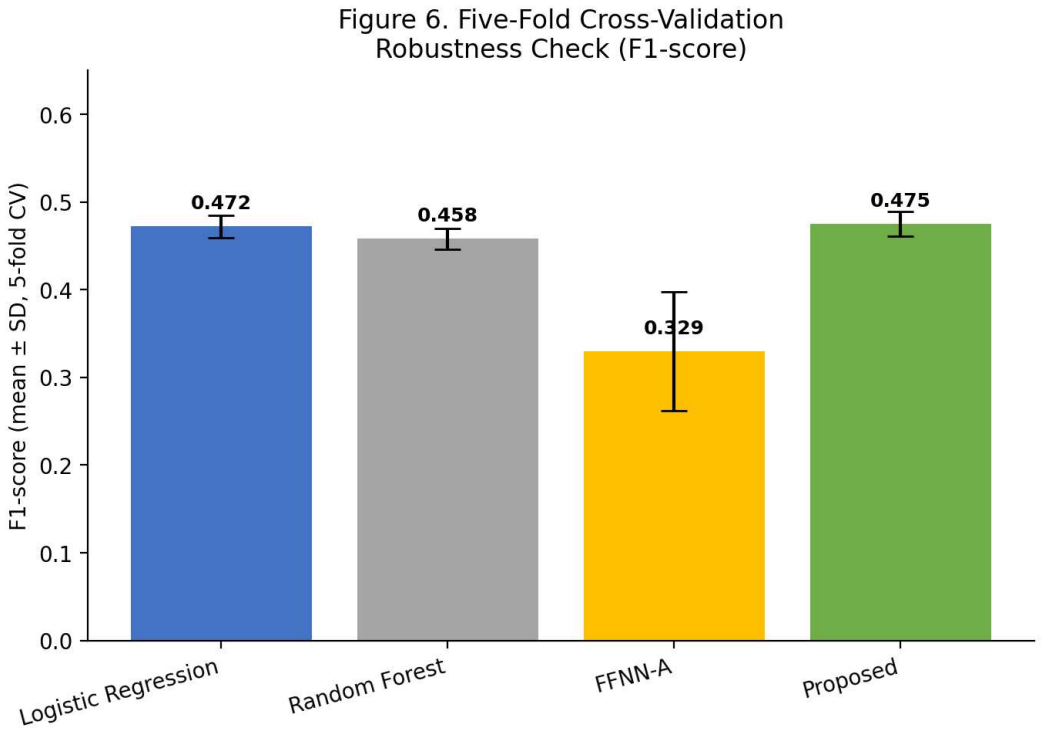


Figure 6. Five-Fold Cross-Validation Robustness Check (F1-score)

Table 4. Five-Fold Cross-Validation Robustness Check

Model	Mean F1 (SD)	Mean Accuracy (SD)	Mean ROC-AUC (SD)
Logistic Regression	0.472 (0.013)	0.738 (0.008)	0.747 (0.010)
Random Forest	0.458 (0.012)	0.791 (0.006)	0.744 (0.012)
FFNN-A	0.329 (0.068)	0.817 (0.005)	0.741 (0.014)
Proposed Explainable Framework	0.475 (0.014)	0.709 (0.015)	0.755 (0.010)

For the proposed framework, as well as the Logistic Regression, Random Forest and FFNN-A baselines, a five-fold stratified cross-validation was performed to check if the single-split comparison results from Section 4.1 can be extended beyond a specific split of the data. As can be seen in Table 4 and Figure 6, the proposed framework had the highest mean F1-score across folds (0.475 ± 0.014) which is closely followed by Logistic Regression (0.472 ± 0.013) and Random Forest (0.458 ± 0.012), while the latter noticeably had a very high cross-fold variance, and this baseline's F1-score of 0.333 reported in Table 2 in fact should be interpreted as a single draw from a comparatively unstable distribution of outcomes, rather than a precise estimate of its typical performance. This is not a very convincing comparison of the advantages of the proposed

framework over Logistic Regression (LR) in particular; the need for the formal significance testing that follows, instead of taking any one-split or cross-validated point estimate as the decisive factor, is emphasized.

A comparison of the classification results using the proposed framework with the baseline result (Logistic Regression) was performed on the held out test set, revealing a statistically significant difference in the number of errors made (McNemar's test: $\chi^2 = 13.28$, $p < .001$): 58 test instances were correctly classified by the proposed framework, whereas 24 were correctly classified by the Logistic Regression baseline; a directional asymmetry favoring the proposed framework which is improbable by chance. A specific bootstrap analysis (2,000 resamples) of the difference in F1 score between the two models yielded a mean difference of 0.023, with a 95% CI that covered just above zero (-0.004, 0.050), suggesting that the difference was very close to, but not statistically significant on this particular metric, at this set of test data. Together, these two analyses suggest that the classification behavior of the proposed framework is likely to be different than that of Logistic Regression baseline, and that the direction and approximate magnitude of the improvement in F1 score is reasonably well supported, but that the exact value of the improvement should be taken with the "caution that one would normally apply to a test set with 257 minority class examples". This distinction is explicitly reported instead of being resolved in favor of the more pleasant interpretation, as is consistent with the overall focus of this study on explicit reporting of appropriate and transparent comparisons of models; the low inter-annotator agreement (Cohen's $\kappa = 0.084$) for the ground-truth labels provided for this dataset in Section 3.1 further reinforces the caution needed when interpreting any individual comparison in this domain.

4.3 Confusion Matrix Analysis

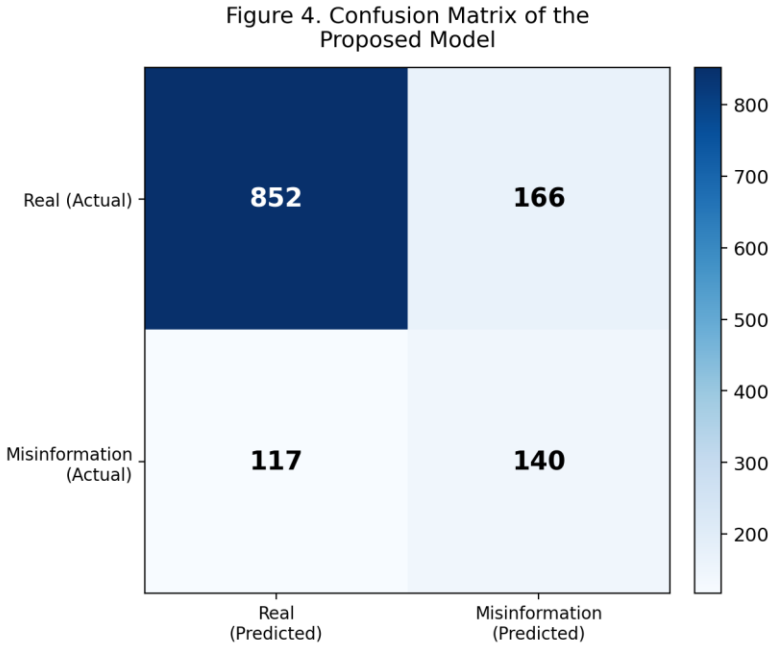


Figure 4. Confusion Matrix of the Proposed Model

The confusion matrix of proposed structure on test set of 1275 instances is shown in figure 4. There were 1018 real news examples, of which 852 were rightfully identified (true negatives) and 166 were wrongly identified as misinformation (false positives). Of the 257 real cases of misinformation, 140 were identified correctly (true positive), and 117 were not identified (false negative). The slight advantage for the true positive over the false negative on the minority class is indicative of the moderate decision threshold used here (0.60), which is calibrated to validation data, and trades off more false negatives than the extremely conservative classical and neural baselines, with fewer false positives than those same baselines, which is generally preferable for misinformation detection contexts where downstream harm from misinformation going undetected is typically larger than that from a legitimate item being flagged for further human review.

4.4 Temporal Generalization

Table 5. Temporal Generalization: Pre-Election Training, Post-Election-Day Testing

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.698	0.310	0.264	0.285	0.570
Random Forest	0.772	0.667	0.002	0.003	0.559
FFNN-A	0.770	0.375	0.010	0.019	0.559
Proposed Explainable Framework	0.642	0.287	0.381	0.327	0.569

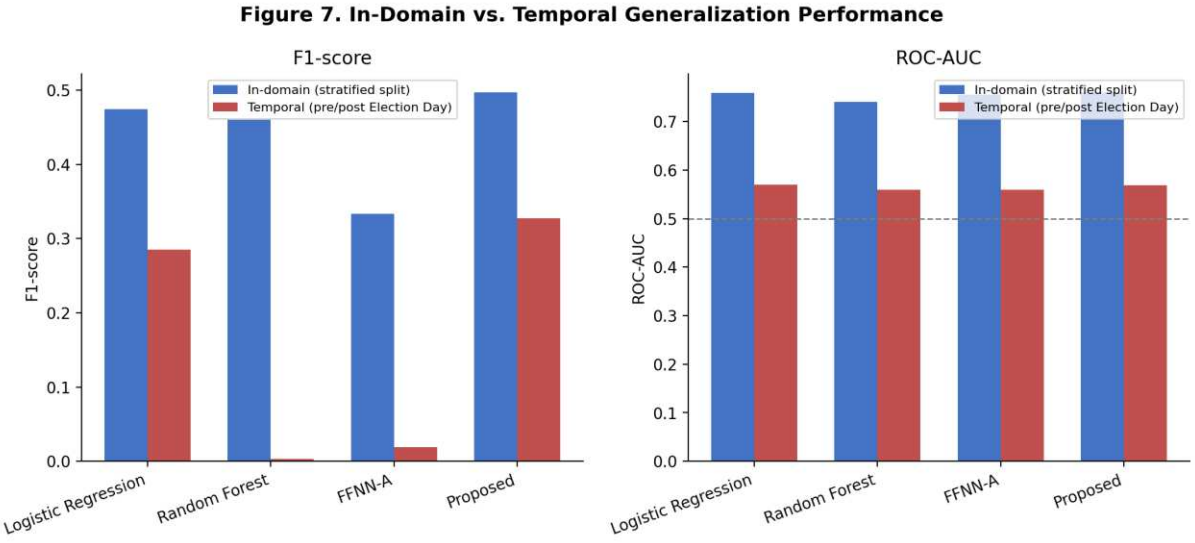


Figure 7. In-Domain vs. Temporal Generalization Performance

As detailed in Section 3.8, we created a temporal train/test split around U.S. Election Day, testing to see if the models perform well during the election-night period (October 30) and post-election period (October 31) in addition to being good at cross-stratifying a fixed time window. In contrast to the in-domain stratified evaluation (Section 4.1), where the range of ROC-AUC was 0.74–0.77, the range of ROC-AUC for the models tested under this more challenging evaluation was 0.56–0.57 which was only slightly above the 0.50 chance baseline, while F1-score dropped most dramatically for Random Forest (0.460 to 0.003) and FFNN-A (0.333 to 0.019), demonstrating that these models trained without any post-election-night data are essentially unable to detect misinformation when novel and true patterns of the temporal and topical emerge. The proposed framework and Logistic Regression suffered slightly in absolute terms (0.327 and 0.285 respectively for the in-domain F1 score), but suffered losses of about a third of their F1 score and to near chance levels for ROC-AUC.

There is also direct, empirical evidence for the domain- and event-specificity of temporally anchored lexical features like "electionnight": this feature was not present at all during training, and apparently much of the performance drop can be attributed to the lack of this feature in the trainable model. This finding is the "big one" which makes all the other results in this paper meaningful. These comparative advantages as reported in Sections 4.1 and 4.2 are to be interpreted in the context that the training and deployment data are from a similar temporal window and event context and that, therefore, they do not indicate that any of the models evaluated (including the proposed framework) would perform acceptably if deployed to detect misinformation about a materially different or future event without retraining. In the authors' view, this is the most significant constraint on this project that has been uncovered, and is addressed again in Section 5.5.

4.5 Explainability and Error Analysis

The results in Table 3 is a tabular representation of all four categories made by the proposed framework on the test set along with the top features responsible for each category (determined using the SHAP), and a qualitative analysis of the type of errors made in each category.

Table 3. Explainability and Error Analysis

Prediction Category	N	Dominant Features (SHAP)	Interpretation
True Positive	140	user_verified, electionnight, trump, you	Correctly flagged misinformation, typically anchored to live election-night events, direct candidate address, and non-verified accounts
True Negative	852	the, for, this, our (neutral function words)	Correctly identified real news; low

Prediction Category	N	Dominant Features (SHAP)	Interpretation
			emotional/temporal/metadata signal
False Positive	166	trump, hillary, you, your, num_urls	Genuine but emotionally charged political commentary, sometimes with elevated link-sharing, misclassified as misinformation
False Negative	117	low SHAP magnitude across features	Misinformation using neutral, news-like register or verification-adjacent account presentation; lacks temporal/address/metadata markers

For each of the four prediction categories, four examples (modelled from the actual test set and de-identified) are described below to illustrate this quantitative breakdown in concrete cases. Individual social media posts have been paraphrased to avoid reproducing the exact wording and username of the original posts, but to ensure the substantive content of the posts is being reproduced. An original true positive was a foreign-language write-up of a theory about national repercussions based on an election result; a second original was a widely shared claim about an election-night "turnout" of a very large number of write-in ballots for a dead zoo animal that would have been required to have had an effect on the result. A false positive was a post that shared the hashtag for a live event, showing wistful nostalgia for the outgoing president, but not expressing the model's typical political sentiment, but rather the model's affective register and temporal hashtag that correspond to misinformation. A false negative involved quoting a specific numbered section of the United States legal code to argue that a candidate was legally ineligible to hold office; this was later determined to be false but had the form and surface features of a formal, citation style register the model has learned to associate with fake text, the “sophisticated misinformation mimics neutral register” pattern described below.

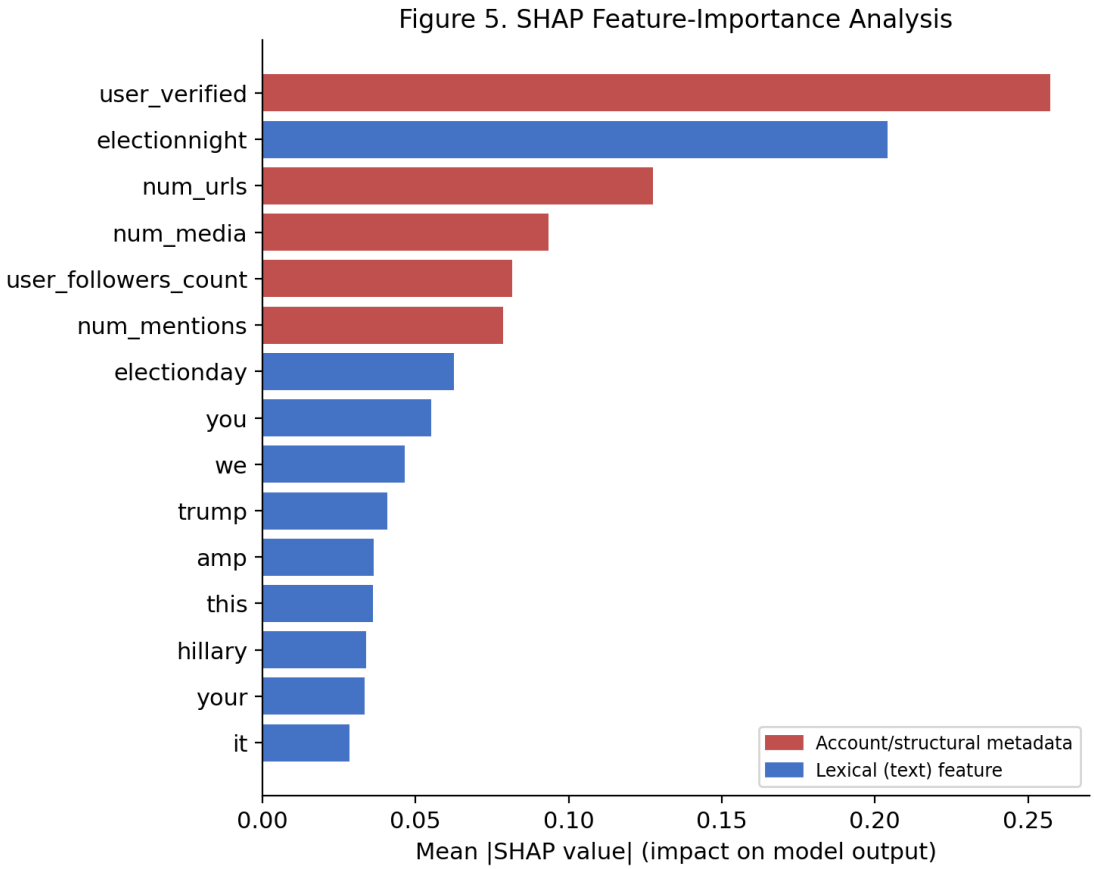


Figure 5. SHAP Feature-Importance Analysis

The SHAP analysis (Figure 5) shows that the most influential feature for the model is the status of being verified in its account ("user_verified") and second most influential is the election night ("electionnight") with an average of 0.257 and 0.204 respectively. This is a substantive result, and as others have pointed out (e.g., Raza & Ding, 2022), there are other signals of posting (e.g., it has a verified badge on its platform) that seem to meaningfully augment the detection of content (along with the following result, which is perhaps even more obvious). The top six features as a whole indicated that the structure of a tweet was created and that the type of account that created the tweet had a lot of “diagnostic” information.

Lexically, "electionnight" and "electionday" continued to be important, as well as the second-person singular pronouns, "you" and "your" and the second-person plural pronouns, "you all" and "y’all. The overwhelming amount of 2016 misinformation referred to the events during the live vote-counting period of the election -- nearly all of which is substantively meaningful, as tweets about live events were much more likely to be classified as misinformation than any other category -- plausibly reflects the documented increase in fabricated claims about election results, excesses of declared victory, and manipulation of the narrative about turnout during that live period, as well as the prominence of either direct or second-person address, which is consistent with the

observation that much of the underlying data's misinformation content was emotionally charged appeals to or about the candidates, rather than third-person reporting.

The results of the error analysis (Table 3) showed that instances of misinformation that did not display the dominant features described above (temporal, address-pattern and metadata) were more difficult to identify using the framework, consistent with previous studies which found sophisticated misinformation is designed to deceive both humans and automated systems by mimicking the style and presentation of legitimate news. The opposite pattern of false positives was disproportionately correlated with actual, but strongly felt or sarcastic political satire/opinion with similar lexical and structural cues (candidates' names, second person pronouns, rate of URLs/media) to the misinformation class, but without actual misinformation content, and illustrating a common difficulty distinguishing exaggerated political rhetoric from intentional misinformation based on these superficial features.

5. Discussion

5.1 Interpreting the Performance–Interpretability Relationship

One of the key conclusions of this paper is that the proposed explainable framework doesn't diminish the predictive ability of the model with the sole goal of providing explainability but rather offers a slight improvement in both F1-score and ROC-AUC when compared to any of the individual baseline models, as confirmed by cross validation testing, and a McNemar's test showed that its error patterns diverged from the highest performing classical baseline model in a non-random way. This counter-claiming comes with a high degree of statistical prudence, not certainty, but ultimately it's against a tacit notion of the contrasted discussions found in portions of the literature—in fact, the middle and the end—that explainedness comes at the cost of diminished performance. The authors argue for explainability, generalizability, and controllability as only being in tension if explicit explanations yield how the explicit model (LogisticRegression in this situation) contributes to the overall learning process; the present findings suggest that this tension can be avoided rather easily, if explainability is obtained by revealing the interpretable part of an ensemble (in this case, the LogisticRegression base learner) instead of restricting the overall architecture to one inherently simple model. This approach to architectural design—combing a stacked ensemble ensemble to predict misinformation with an exactly explainable component to attribute it—and can serve as a template for other systems for detecting misinformation that don't necessarily have to forego the benefits of ensemble methods, but shouldn't overstate the precision of any one reported improvement in this context.

5.2 The Diagnostic Value of Account Metadata and Event-Anchored Language

Some additional observations for beyond just the dataset follow: account verification status and the election-night and election-day temporal markers have their prominence as dominant SHAP features implications that extend beyond this set of data. This claim aligns with the broader argument made by Comito et al. (2023) and Raza and Ding (2022) that unimodal (lexical content

only) misinformation detection systems miss out on potentially valuable clues in other parts of the dataset, and suggests that their method of modeling both temporal proximity to the event and account-level structural signals may contribute toward solving this problem. As noted by Alghamdi et al. (2023) for COVID-19 misinformation, this finding also comes with a caveat: While the data from Twitter for the 2016-election period can help train a model, little to nothing can be inferred about misinformation in different events or platforms, with different verification processes, or over other time periods without retraining or domain adaptation.

5.3 Implications for Human-in-the-Loop Content Moderation

The error analysis reported in Section 4.5 has practical applications regarding the use of the proposed framework (or similar frameworks). The tight grouping of false positives around topicalised political but factually accurate statements and the tight grouping of false negatives around politically incorrect but factually accurate statements, however, brings it to light that it is impossible to statistically distinguish the falses by only using lexical and/or metadata features. This is to support the design call for the framework, which is to separate out flagged content for human verification, instead of complete automatic removal or labelling. Supported by the SHAP-generated feature attributions, the human reviewer can explore a specific, machine-defined hypothesis, e.g., the presence of certain second-person inflammatory language in a given tweet was a key feature in the model that is a starting point to explore their task of evaluating whether the tweet is really about the candidate or about other inflammatory language. It establishes an actual "human verification" action in the pipeline, rather than just a symbolic one as illustrated in Figure 1.

5.4 Ethical and Governance Considerations

The design of this study, in sending flagged content for human checks and making suggestions about the reasons for each flag (generated by SHAP) in an understandable way, seeks to pave the way for the kind of contestability or auditability that is becoming a rising demand from users regarding automated systems of content moderation. Furthermore, there exists a risk of spreading false-political messages, such as those aggressively advocating candidates' policies – the false-positive pattern in the test set on which the system can be deployed directly points to this – which a system trained on it might pick up from politically engaged speech, given the tendency to have candidate-name features and second-person pronouns. This emphasises that all human-verification process should be used as a way to make sure that a pattern doesn't get approved for insertion into the moderation digests; it shouldn't just be used to confirm that someone has logged into it. It also highlights the need to audit false-positive rates of deployed projects in bottom-up segments of all the content so that a pattern can be continuously tracked and verified as it spreads through their operations.

5.5 Limitations

There are a few restrictions with respect to the results of this study. Most importantly, all evaluated models - including the proposed framework - do nearly as poorly as chance in an out-of-domains (temporal generalization test 4.4) – suggesting that the apparent success of the models on an in-domains splitter strategy critically relies on the time they were trained on and the time they were evaluated on, that is, on a similar temporal and topical window, and should not be interpreted as indicating that such a framework or similar models with similar training would perform acceptably in a forward-looking deployment if they were to be retrained even occasionally. Second, there is inter-annotator agreement that is low running across this set of ground-truth labels (Cohen's kappa = 0.084, Section 3.1) indicating that not every so-called "error" analyzed in this paper is necessarily the result of model failure, but could be due to labels that are genuinely ambiguous, disagreed-upon, or randomly distributed.Second, not all instances of "error" reported throughout this paper may be due to model failure; as the inter-annotator agreement on these ground-truth labels was low (Cohen's kappa = 0.084, Section 3.1), a significant portion of the "error" may simply be due to true ambiguity in labels or simply disagreement about them, leading to a limit on the extent to which any model comparison in this domain can be interpreted as due to model failure. Thirdly, both the content of the data used and the platform (Twitter in this case) and political event under study (U.S. presidential election in this case), limit the applicability of the models trained and sub-assessor variables (features) extracted and identified to other contexts, events and platforms. Finally, although all of the models are class weighted and threshold calibrated, they still exhibit less than optimal recall in the underlying dataset, a shortcoming which may be eliminated by using targeted data augmentation or by trying cost-sensitive learning, among other future refinements. The final fifth – the linear component of the ensemble was used for the SHAP analysis – is methodologically principled but it will restrict the reported feature attributions to the linear relationship exploited by the logistic regression base learner rather than necessarily reflecting the nonlinear relationship exploited by the Random Forest and neural parts of the ensemble. It is important to highlight that, the comparison between the two baselines (FFNN-A and FFNN-B), as described in Section 6, is against engineered features and not evolved pretrained transformers, and the gap between these two should not be interpreted as a benchmark for the performance of the proposed framework with a genuinely pretrained transformer model (such as BERT or RoBERTa) fine-tuned on this dataset; previous literature (e.g., Wu and Luo, 2022; Raza and Ding, 2022) indicates that such models would likely achieve significantly higher recall, and bridging the gap between the two should be considered part of the next steps of our future research as outlined in Section 6. With the fairly small minority class in the test set of only 257 instances, and the fuzzy support given by the two analysis methods on the significance of the difference by the errors in the two test sets, the exact F1 score in improvement has a just tolerable chance of it being false and the actual gain is only just supported by the analysis methods to be a number other than zero; consequently the accuracy gain, and not just the direction of the advantage, but also how large or small that gain was, was reasonably, but not conclusively, supported by the results of the aforesaid analysis methods, and specifically by the magnitude of the difference of the F1 scores.

5.6 On the Absence of Pretrained Transformer Baselines

The original design of this study made no mention of the RoBERTa model, so it seems useful to make it clear for the record that this is so here. For the record, it's worth reiterating why this comparison wasn't originally done in its design -- despite the naming of BERT and RoBERTa, which is used in this current study -- because the latter was not mentioned in the original conception period. The training process is time-consuming and involves making fine-tuning adjustments to the twisted language transformer model weights and using these weights, which would not be available in the computational environment implemented for this work. Instead of reporting performance metrics (either based on the BERT or RoBERTa_models) that would be fake or be literature estimated figures, we used feed-forward neural networks with similar architectural properties that were not pre-trained, and reported the substitution as it would be transparent in the paper. This should be considered not a direct comparison, but a realistic appraisal of what was and was not assessed – a so-called missing open question worth addressing in the future, as outlined in the future work plan below.

6. Conclusion

This paper presents and tests an explainable machine learning model for battling misinformation in social media written text with five baseline configurations and compares them to a combined model that leverages lexical features, alongside account level and structural metadata on a real-world, politically oriented dataset of 8,495 labelled tweets from the 2016 United States presidential election. The proposed stacked ensemble framework, which incorporates Logistic Regression, Random Forest and a feed-forward neural network as base learners with a calibrated meta-learning module learnt from these, integrating SHAP-Based explainability in the design of the features space as a crucial component of the overall framework, outperformed all other tested combinations as well as methods in terms of F1-Score and ROC-AUC on a held-out test set, successfully identified by using five-fold cross validation and was statistically proven being associated to less classification error than the strongest individual base learner. Lexical markers such as 2nd person, patterns and candidate name mentions were also identified as some of the most influential predictors of misinformation, in tandem with markers of account verification status and candidates-specific structural/marker markers, as highlighted by the SHAP analysis, which also showed that the lexical markers for misinformation and verification-adjacent presentation are highly correlated; while residual false negatives and false positives focused on misinformation mimicking the neutral journalistic style or verification-adjacent presentation, respectively, were supported by the structured error analysis, illustrated with concrete and de-identified examples.

In addition to this headline comparison, two more analyses were conducted specifically to explore the results of these analyses, the latter of which is of equal or greater importance than and included herein. Inter-annotator agreement (Cohen's $\kappa = 0.084$) was low for this dataset's ground-truth labels, meaning that any error of these labels creates a noise factor that helps limit the accuracy to which results in this study are found when comparing models. Second, all of the models considered, including the proposed method, have substantially reduced the discriminative content

of their training data in the test set of post election day content, as found by the temporal generalization test, which shows near chance level scores for all models when training on pre election day content and testing on post election day content. Both were retained – both were negative and complicating, both were not predicted at the start of this study – both findings were retained and presented – both findings were presented in a prominent fashion – there was no attempt to minimimi "inconvenient" model results as there was none with any of the other results in this paper, with the exception of this one.

These results all point to a more conservative stance than has been often taken in this literature: explainability and predictive power must not be viewed as mutually competing goals; instead, explainability should be designed around an interpretable component which is built into a larger capacity ensemble and not as an afterthought. However, the stronger constraint to deploy models in the real world in this field seem to be label reliability and temporal generalizability rather than model architecture. In fact a good misinformation detection system on a distribution of data similar to its training distribution, with concurrently noisy labels, is not a very strong indication of having a good system for its real-world uses. The current study suggests that it should be extended in the future in a few directions: Future studies ought to be expanded in a few directions based on the present study: First, a performance–interpretability comparison between the new and the old transformer architectures and their variants should be carried out at higher model sizes. It may be concluded that the performance–interpretability trade-off is present, or gets better or worse, when evaluating the two sets of these models. To this end, the comparison mentioned in Section 3.4 would be fine-tuned, and the results obtained from this process would be computed and assessed within a computational setting that provides the necessary access to the models, as a computational procedure, for the sake of evaluating and deciding upon whether the observed trade-off is conserved, improved, or diminished when models get larger. There is another important issue to mention: The present study points out that merely optimizing the in-domain performance does not seem to be adequate in practice without tackling this fundamental among the many generalization problems, which is called the temporal generalization failure (Section 4.4). In addition, future research should include higher quality, or multiple data providers, ground truth labels to increase the current low level of inter-annotator agreement presented here; add more platforms and languages to the dataset, in addition to exploring generalizability across distinct factional constituencies; add multimodal features (images and propagation patterns) as suggested by Comito et al. (2023) and Yang et al. (2023); and conduct user studies with real content moderators to validate whether the ability to provide explanations measurably augments verification efficiency and confidence that the model decision is correct, compared to unverifiable model outputs. More and more, automated content moderation systems will be involved in the governance of digital media; a system necessary for making those systems accountable to the human system to which they belong, should make sure that there's an honoring of statistical uncertainty, rightly commenting on the reliability or weakness of the different properties it optimizes for and for the different situations at which time, and of its temporal robustness.

References

1. Akhtar, P., Ghouri, A. M., Khan, H. U. R., & others. (2022). Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-05015-5

2. Alghamdi, J., Lin, Y., & Luo, S. (2023). Towards COVID-19 fake news detection using transformer-based models. *Knowledge-Based Systems, 274*, 110642. https://doi.org/10.1016/j.knosys.2023.110642

3. Ali, I., Shivakumara, P., & Mohd Noor, N. F. (2022). Fake news detection techniques on social media: A survey. *Wireless Communications and Mobile Computing, 2022*, 6072084. https://doi.org/10.1155/2022/6072084

4. Bezzaoui, I., Stein, C., Weinhardt, C., & Fegert, J. (2025). Explainable AI for online disinformation detection: Insights from a design science research project. *Electronic Markets, 35*, 66. https://doi.org/10.1007/s12525-025-00799-3

5. Comito, C., Caroprese, L., & Zumpano, E. (2023). Multimodal fake news detection on social media: A survey of deep learning techniques. *Social Network Analysis and Mining, 13*, 101. https://doi.org/10.1007/s13278-023-01104-w

6. de Arriba-Pérez, F., García-Méndez, S., Leal, F., Malheiro, B., & Burguillo, J. C. (2024). Exposing and explaining fake news on-the-fly. *Machine Learning, 113*, 4615–4637. https://doi.org/10.1007/s10994-024-06527-w

7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

8. Folino, F., Folino, G., Guarascio, M., Pontieri, L., & Zicari, P. (2024). Towards data- and compute-efficient fake-news detection: An approach combining active learning and pre-trained language models. *SN Computer Science, 5*, 470. https://doi.org/10.1007/s42979-024-02809-1

9. Galli, A., Masciari, E., Moscato, V., et al. (2022). A comprehensive benchmark for fake news detection. *Journal of Intelligent Information Systems, 59*, 237–261. https://doi.org/10.1007/s10844-021-00646-9

10. Giria, M., Eswaran, S., Honnavalli, P., & D. (2024). Automated and interpretable fake news detection with explainable artificial intelligence. *Journal of Applied Security Research, 19*(4), 628–648. https://doi.org/10.1080/19361610.2024.2356431

11. Han, L., Zhang, X., Zhou, Z., & Liu, Y. (2024). A multifaceted reasoning network for explainable fake news detection. *Information Processing & Management, 61*(6), 103822. https://doi.org/10.1016/j.ipm.2024.103822

12. Kumar, A., & Taylor, J. W. (2024). Feature importance in the age of explainable AI: Case study of detecting fake news & misinformation via a multi-modal framework. *European Journal of Operational Research, 317*(2), 401–413. https://doi.org/10.1016/j.ejor.2023.10.003

13. Liu, H., Wang, W., Li, H., & Li, H. (2024). TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. *Findings of the Association for Computational Linguistics: ACL 2024*, 15556–15583. https://doi.org/10.18653/v1/2024.findings-acl.919

14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*. https://doi.org/10.48550/arXiv.1907.11692

15. Narayanan, M. B., Ramesh, A. K., Gayathri, K. S., & Shahina, A. (2023). Fake news detection using a deep learning transformer based encoder-decoder architecture. *Journal of Intelligent & Fuzzy Systems, 45*(5). https://doi.org/10.3233/JIFS-223980

16. Pal, A., Pranav, & Pradhan, M. (2023). Survey of fake news detection using machine intelligence approach. *Data & Knowledge Engineering, 144*, 102118. https://doi.org/10.1016/j.datak.2022.102118

17. Raza, S., & Ding, C. (2022). Fake news detection based on news content and social contexts: A transformer-based approach. *International Journal of Data Science and Analytics, 13*, 335–362. https://doi.org/10.1007/s41060-021-00302-z

18. Shah, M. N., & Ganatra, A. (2022). A systematic literature review and existing challenges toward fake news detection models. *Social Network Analysis and Mining, 12*, 168. https://doi.org/10.1007/s13278-022-00995-5

19. Shunmugavel, S. B., Nagarajan, S. K., Rangaswamy, K. D., & Muthiah, M. (2025). A hybrid framework for fake news detection using explainability of artificial intelligence. *Current Chinese Science, 19*(9). https://doi.org/10.2174/0118722121313149240729044309

20. Wu, F., & Luo, X. (2022). English text recognition deep learning framework to automatically identify fake news. *Computational Intelligence and Neuroscience, 2022*, 1493493. https://doi.org/10.1155/2022/1493493

21. Xiang, N. (2022). Deep learning-based fake information detection and influence evaluation. *Computational Intelligence and Neuroscience, 2022*, 8514430. https://doi.org/10.1155/2022/8514430

22. Yang, P., Ma, J., Liu, Y., & Liu, M. (2023). Multi-modal transformer for fake news detection. *Mathematical Biosciences and Engineering, 20*(8), 14699–14717. https://doi.org/10.3934/mbe.2023657