

Original Article

Comparative Analysis of Machine Learning Models with SHAP Explanations towards Transparent Fake News Detection

Ramoni Tirimisiyu Amosa¹, Adekiigbe Adebajo², Fabiyi Aderanti Alifat², Olorunlomeye Adam Biodun³, Adisa Solagbade Philip⁴, Lawal Moshood Olatunji⁵, Akanni Abideen Amosa⁶

¹Department of Computer Science, Federal Polytechnic Ede, Osun State, Nigeria.

²Department of Networking & Cloud Computing, Federal Polytechnic Ede, Osun State, Nigeria.

³Department of Intelligent Systems Engineering, Obafemi Awolowo University, Ile Ife, Osun State, Nigeria.

⁴ICT Unit, Federal Polytechnic Ede, Osun State, Nigeria.

⁵Department of Cyber Security, Federal Polytechnic Ede, Osun State, Nigeria.

⁶Department of Computer Engineering, Federal Polytechnic Ede, Osun State, Nigeria.

¹Corresponding Author : ramoni.tirimisiyu@federalpolyede.edu.ng

Received: 16 July 2026

Revised: 21 August 2026

Accepted: 08 September 2026

Published: 29 September 2026

Abstract - The generation of fake news on online platforms has grown into a serious problem, compromising public trust, public opinion and the possibility of information-based decision making. In this study, an intelligent Fake News Detection (FND) system to automatically classify news articles as fake or real was developed using techniques of Natural Language Processing (NLP), Machine Learning (ML) and Explainable Artificial Intelligence (XAI). A publicly available dataset consisting of 44,856 news articles from the LIAR data set and the Kaggle repository was used, of which 80% was used for training and 20% for testing. The data preprocessing included the following steps: Removing punctuation, Lowercase, Stop-word removal, Tokenization, Stemming, Lemmatization, and TF-IDF (Term Frequency – Inverse Document Frequency) Feature Extraction. The classification algorithms used were Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Long Short-Term Memory (LSTM), which were trained and evaluated with accuracy, precision, recall, F1-score, and AUC-ROC. Experimental results indicated that SVM outperforms the other models with respect to accuracy (99.21%), precision (99.25%), recall (99.23%), F1-Score (99.24%) and AUC-ROC (99.96%). To make the classifications auditable, SHapley Additive exPlanations (SHAP) was applied uniformly to all five trained models rather than to the best model alone, permitting a direct comparison of the lexical evidence on which each classifier relies. The attributions converge: four of the five models rank the token “said” as the single most influential feature, and in every case, it shifts the prediction towards REAL, indicating that the models are detecting the attribution conventions of professional reporting rather than topic alone. The results show that the proposed NLP-ML-XAI framework is accurate, interpretable, and reliable for automatic fake news detection, which could be deployed by media organizations, fact checkers, and digital platforms to prevent spreading fake news online.

Keywords - Dataset, Fake News, Machine Learning, Prediction, Repository.

1. Introduction

Miscommunication through the mass dissemination of fake news is a serious issue in the current digitalized world, which can have many consequences on the personal, institutional and civilisational level. The information that is rapidly spreading hampers democratic processes, informed decision-making, and the credibility of authoritative information (Padalko et al., 2023). With the plethora of communication technologies and media platforms, everyone can access digital news, which aids in spreading fake news. Fake news is the dissemination of False or Falsified news on social media platforms, which causes harm to an individual, organisation and/or governmental body (Malhotra and Malik, 2024; Akter et al., 2025). Fake news, generally speaking, is false or misleading content, of the “journalistic” variety,

designed to mislead readers. News media that is false, but is organized differently from or has a different purpose from that of news media, and is similar in form to news media. Fake news is created to “manipulate” people’s opinions, to impress a person, or to make money from online advertisements. According to Priya et al. (2026), fake news is news reports that are knowingly and explicitly false, and may mislead readers. Alghamdi et al. (2018) describe fake news as published content that resembles a product from news media, can be disseminated rapidly on the internet, usually through social media networks. Academics are worried about the propagation of fake information due to its potentially severe implications on democratic processes and information system trust. As these definitions remind us, fake news is not just a matter of misinformation, but also often spread and fabricated for the sole intention of being inaccurate.



Fake news may not be so much or so little, may be more or less shaped, to its use, to its presentation and to its dissemination (Rashkin et al., 2017). Researchers further noted that there are three deceptive tactics: satire or parody, propaganda, misinformation and disinformation. The purpose of satire/parody is usually to entertain or amuse, but some readers may take it as fact. Propaganda: information that is intentionally misrepresented in some way to instill a certain attitude or agenda in people. A person may give a person on the street misinformation that is good-hearted, but without purposeful misinformation, and may not be sure whether it is true or not. On the other hand, disinformation is intentionally created and spreads misinformation to deceive, mislead or confuse the truth. The various types of fake news are a reminder that disinformation is a multifaceted phenomenon, and that there are methods other than words for determining a fake, and the fake news agent has the ability to contact numerous users via social media like WhatsApp, Twitter and Facebook.

While there is lots of good stuff about this interconnectedness, it can also create a situation in which false information can fly around before it can be verified or debunked. According to Zhou & Zafarani (2020), the users of social media platforms tend to post sensational or emotion-triggering information more than accurate information, which eventually leads to propagating bogus information. Also, the algorithms used by social media platforms often prioritize content that has a high level of engagement, such as likes, comments, and shares. This can be indirect spreading of misinformation that can have an impact on public opinion. If news content is not verified, then fake news will quickly spread in the Online Social Networks (OSNs) and would have a great impact. Therefore, the researchers interested in NLP would like to apply a number of ML and DL models to identify fake news. The increasing volume of websites, blogs, social media and news articles that are beginning to generate digital text data has made NLP techniques more applicable to automatically analysing and deriving information from text. NLP plays a crucial role in the field of text analysis, particularly given its ability to process and comprehend human speech. Because most high-performing machine learning and deep learning classifiers operate as black-box systems, the classifier return a fake/real verdict without any account of why the verdict was reached, without an explanation of a label, a fact-checking desk can do nothing, an editor can do nothing in the case of a takedown, and a regulator cannot do anything to audit a system with an unknown basis behind the decisions. Explainable Artificial Intelligence (XAI) tackles this by explaining the reasoning behind a prediction based on the specific features in the input to the AI (Gongane et al., 2024; Munawaroh, 2024; Bezzaoui et al., 2025).

Three limitations recur in the existing Fake News Detection (FND) literature. First, most comparative studies benchmark accuracy only, and report the winning algorithm without examining the linguistic basis of the decision; the reader is therefore unable to judge whether high accuracy reflects genuine deception cues or dataset-specific artefacts. Second, studies that do report explanations typically explain

a single model, so it is not known whether different classifiers that agree on the label also agree on the evidence. Third, comparative benchmarks are frequently reported without full disclosure of preprocessing, class balance, split protocol and hyperparameter settings, which makes the reported figures difficult to reproduce and difficult to compare across papers.

Alghamdi et al. (2022), for example, provide an extensive multi-dataset benchmark of classical, deep and transformer models, but explicitly leave the interpretability of the outputs as an open limitation. Rathore et al. (2024) and Neela et al. (2024) report accuracies above 97% with TF-IDF features but offer no feature-attribution analysis. Alghamdi (2024) and Alghamdi (2024) obtain strong deep and transformer results at the cost of computational load and cross-domain fragility, the gap addressed in this study is therefore the absence of a single, reproducible study that (i) benchmarks classical and deep learning classifiers on the same corpus under identical preprocessing and splits, and (ii) applies a common feature-attribution method to all of the benchmarked models so that their decision bases can be compared directly, not merely their scores.

2. Literature Review

2.1. ML, DL Approaches to FND and Feature Representation for Textual Deception Detection

With the emergence of fake news as a problem, there are several automatic systems that have been developed to detect fake news using ML and NLP techniques. There have been a number of papers that have concentrated on the creation of various algorithms and datasets, in addition to feature extraction techniques that would enhance the accuracy of fake news classification. A few research projects are considered in this section, focusing on the methods used, the databases employed, the results obtained, and the limitations highlighted. Traditionally, newspapers and TV channels are the best sources of news and information for people. The easy access has made OSNs much more important. As a result, it is now more than just a means of communication between people; rather, it is a crucial instrument for information sharing as well as influencing and forming public opinion (Kwak et al., 2010). The proliferation of fake news, particularly on OSNs, is, however, a serious problem for both individuals and society at large because of its lack of supervision, control and automated fact-checking, and the production of low-quality and counterfeit information.

This leads to the public being exposed to misinformation and information with falsehoods on OSNs, including fake news (i.e., news articles with deliberate falsehoods to mislead readers) (Allcott & Gentzkow, 2017; Shu et al., 2017). Lazer et al (2018) went on to investigate the spread of fake news on social media platforms and reached the conclusion that fake news is more likely to be spread and more easily diffused than true news, and therefore, automated systems should be developed. Detection performance depends as much on the representation of the text as on the classifier. Three families dominate. Frequency-based representations, chiefly Bag-of-Words and TF-IDF, weight terms by their local frequency against their corpus-wide rarity are sparse, high-dimensional, computationally light and, importantly for this study, directly

interpretable because each dimension corresponds to a single token. Static distributed representations such as Word2Vec and GloVe map tokens into dense vectors that encode distributional similarity but assign one vector per token irrespective of context. Contextual representations produced by transformer encoders assign token vectors conditioned on the surrounding sentence, capturing polysemy and longer dependencies at substantially greater computational cost. Alghamdi et al. (2022) compared representations across all three families and found no uniformly dominant option, with frequency-based features remaining competitive on smaller and mid-sized corpora. Because the present study requires token-level attributions that a domain expert can read, TF-IDF was retained as the representation for the four classical models, and a learned embedding layer was used for the LSTM.

A system that uses NLP techniques to detect bogus news automatically is presented by Rathore et al. (2024). It adopts two types of NLP methods: a semantic method based on an algorithm detecting whether there are inconsistencies in the news articles, and a text-based method based on word embedding. The semantic approach is for recognising textual inconsistencies; the text-based approach is to embed articles in a vector space based on words. The approach successfully identified and classified fake news with good accuracy, offering potential applications for various stakeholders in the fight against misinformation on open platforms. The study evaluates some classification models like SVM, Random Forest Classifier (RFC) and LR. TF-IDF feature extraction gives the highest accuracy of 99.13% along with SVM. The accuracy of the Random Forest model is 87.14%, and the second is the Logistic Regression model with an accuracy of 97.90%. In Neela et al. (2024), NLP techniques were used to process the dataset, which was acquired from Kaggle. Text cleaning, tokenization, removal of stop words and TF-IDF vectorization were the pre-processing steps performed. Gradient Boosting, RF, DT and LR were the algorithms employed. Ensemble majority voting was used to boost the prediction accuracy.

The results were satisfactory, with the best result obtained by Random Forest. The logistic regression model was almost 97.87% accurate. Alghamdi et al. (2022) carried out a detailed benchmarking study of deep learning and machine learning on fake news detection, but the researchers failed to incorporate explainable AI to determine the most influential feature. The researchers evaluated four popular datasets: LIAR, PolitiFact, GossipCop, and COVID-19 using a variety of traditional machine learning algorithms, such as Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, XGBoost and ensemble voting techniques, as well as sophisticated deep learning models including CNN, BiLSTM, BiGRU, CNN-BiLSTM, CNN-BiGRU, and hybrid architectures. Researchers also explored the transformer models such as BERTbase, RoBERTabase, with different feature extraction techniques like TF-IDF, CountVectorizer, Word2Vec, GloVe and contextual embeddings. The study showed that no single transformer model had the best performance with each of the datasets; models tend to perform best when trained with one algorithm or classifier. For

instance, ensemble machine learning systems achieved 63.9% accuracy on the LIAR dataset, and BERT-based systems achieved a performance greater than 98% on the COVID-19 dataset. Results also indicated that traditional machine learning models are competitive on smaller data sets, but on larger data sets, deep learning models outperform the traditional machine learning models. While it is important, it can be acknowledged that the study has several limitations, such as only considering text, excluding images, videos and signals from social networks, the computational complexity of the transformer models, the interpretability of the outputs of deep learning models and the binary classification of FND.

Ahmed et al (2021) employed supervised learning algorithms Random Forest (RF), TF-IDF feature extraction, Naïve Bayes and Support Vector Machine (SVM). The answer of these models was that SVM gave the maximum accuracy, despite having noisy and incomplete datasets, as the reason for this low accuracy in the models. Alghamdi (2024) studied CNN and LSTM deep-learning models and concluded that these models are superior to the traditional models, but were very resource-intensive. Similarly, Alghamdi (2024) used transformer-based models (e.g., BERT) to achieve good FND performance in a multi-lingual setting, whose performance was impacted when the domains or the text style were outside of the training set. Using machine learning and deep learning technology in fake news identification is said to aid, but there are challenges regarding data set quality, computational costs, domain adaptability and the problem of distinguishing satire from fake news.

2.2. Explainable Artificial Intelligence in Fake News Detection

The interpretability requirement in FND is both operational and ethical. Post-hoc, model-agnostic attribution methods are the dominant response. LIME approximates the model locally with an interpretable surrogate fitted on perturbed samples, which is fast but yields explanations that can vary between runs. SHAP derives feature attributions from cooperative game theory, distributing the difference between a prediction and the base value across the input features; the resulting values are additive and satisfy consistency and local accuracy properties, which is why SHAP was adopted here. Attention-weight inspection is available for neural architectures but is contested as an explanation, since attention mass does not reliably indicate feature importance.

Most applications report only a single model, as reported by Gongane et al. (2024), who survey XAI methods for fake news and hate speech. Alike results can be found in Munawaroh (2024) for deep misinformation classifiers, where interpretability is typically regarded as an addendum to the report, not as an element to compare. This study, conducted by Bezzaoui et al. (2025), demonstrates that the quality of the explanation is a significant factor in practitioners' willingness to trust a disinformation detection system. Read together, these findings justify the design used in this study: one single attribution method should be used for each model that is benchmarked, and the agreement or disagreement of the explanations should be considered as an end in itself.

Table 1. Summary of the Literature Review

Study	Dataset	Models compared	Best reported result	Explainability	Gap addressed here
Alghamdi et al. (2022)	LIAR, PolitiFact, GossipCop, COVID-19	Classical, deep and transformer families	63.9% (LIAR); >98% (COVID-19)	None; stated as a limitation	No feature attribution; text-only, binary
Rathore et al. (2024)	Kaggle	SVM, RF, LR	99.13% (SVM + TF-IDF)	None	The decision basis was not examined
Neela et al. (2024)	Kaggle	GB, RF, DT, LR + majority voting	~97.87% (LR); RF best	None	Ensemble gain unexplained
Ahmed et al. (2021)	Multiple	RF, Naïve Bayes, SVM	SVM highest	None	Noise sensitivity; no attribution
Alghamdi (2024)	Large corpora	CNN, LSTM	Deep > classical	None	High compute cost; opaque
Alghamdi (2024)	Multilingual	BERT and variants	Strong in-domain	None	Cross-domain degradation
Bezzaoui (2025)	44,856 articles (Kaggle/LIAR benchmark)	LR, SVM, RF, DT, LSTM under an identical pipeline	99.21% accuracy, 0.9996 AUC-ROC (SVM)	SHAP applied to all five models	Uniform cross-model explanation, convergence finding, full protocol disclosure

3. Methodology

3.1. Research Design

This includes developing and assessing ML models on FND. The idea is to train algorithms on a labeled news corpus and test them on an unseen corpus to correctly classify news articles. The system is built as a text classification model in which the input data (News articles) is processed and transformed into numerical features, which are used by machine learning algorithms for prediction.

3.2. Source of the Dataset

Fake news datasets were taken from a public data repository (kagle.com), which has been used in FND research due to the availability of labeled fake and real news examples to train and test the models consistently across different research. The LIAR dataset, introduced by William Yang Wang (2017), is one of the most commonly used benchmark datasets in FND research. It contains manually labeled political statements collected from the fact-checking website PolitiFact, providing verified ground truth labels for machine learning models. In addition, Kaggle hosts multiple fake news datasets that include full-length news articles and headlines labeled as real or fake, making them suitable for text classification experiments and model comparison. For clarity, the experiments reported in this paper were conducted on the Kaggle ISOT-style corpus, comprising Fake.csv (23,481 records) and True.csv (21,417 records), giving 44,856 usable records after cleaning. The LIAR dataset is discussed as a reference benchmark for contextualising published results on short political statements and was not merged with the Kaggle corpus for training, since the two differ in unit of analysis (full articles versus single statements) and in label granularity (binary versus six-way). Reporting the two separately avoids the label-mapping ambiguity that would otherwise be

introduced by collapsing LIAR's six-point truthfulness scale onto a binary target.

3.3. Data Description

Textual news data with associated labels indicating whether the content is real or fake were used. Each record in the dataset typically includes the following fields:

- News title or headline
- Full news article text
- Label indicating authenticity (e.g., *fake* or *real*)

Such labeled textual datasets are essential for supervised machine learning, where models learn to classify news articles based on patterns present in the training data. In the case of the LIAR dataset, the data includes short political statements along with truthfulness ratings ranging from "true" to "pants on fire," which can be simplified into binary labels for FND tasks. The LIAR dataset contains approximately 12,800 manually labeled statements, making it a widely used benchmark for training and evaluating FND models. Table 1 shows the composition and description of the dataset.

3.4. Sampling and Partitioning Protocol

The full cleaned corpus was used rather than a drawn sample, so no sampling loss was incurred. Partitioning was a single stratified hold-out split of 80% training (35,884 records) and 20% test (8,972 records), stratified on the class label so that the 52.3% fake / 47.7% real balance of the corpus is preserved in both partitions. A fixed random seed (42) was used so that the partition is reproducible. The class distribution is close to balanced, so no resampling, class weighting or synthetic oversampling was applied, and unweighted accuracy is therefore a defensible headline metric; AUC-ROC is nevertheless reported alongside it to

guard against threshold-dependent conclusions. Hyperparameter selection was performed by five-fold stratified cross-validation within the training partition only; the test partition was used once, for the final evaluation reported in Table 4, and at no point were model or parameter selection informed.

Table 2 reports the search space explored for each model and the configuration selected by cross-validated F1-score. Model selection favoured the simplest configuration whose cross-validated performance was statistically indistinguishable from the best, in order to limit overfitting to the corpus.

Table 2. Search space and selected hyperparameters

Model	Search space explored	Selected configuration	Selection basis
LR	$C \in \{0.01, 0.1, 1, 10, 100\}$; penalty $\in \{L1, L2\}$; solver $\in \{\text{liblinear}, \text{lbfgs}\}$	$C = 10$, L2 penalty, liblinear, max_iter = 1000	5-fold CV F1
SVM	Linear kernel; $C \in \{0.1, 1, 10\}$; RBF with $\gamma \in \{\text{scale}, 0.01\}$	LinearSVC, $C = 1$, squared hinge loss; probabilities via Platt calibration for AUC-ROC	5-fold CV F1; linear kernel selected as RBF gave no significant gain at higher cost
RF	n_estimators $\in \{100, 150, 300\}$; max_depth $\in \{\text{None}, 30, 60\}$; min_samples_leaf $\in \{1, 2, 5\}$	150 trees, max_depth = None, min_samples_leaf = 1, Gini criterion	5-fold CV F1; 300 trees gave <0.1% gain
Decision Tree	max_depth $\in \{10, 30, \text{None}\}$; criterion $\in \{\text{Gini}, \text{entropy}\}$; min_samples_split $\in \{2, 10\}$	Gini, max_depth = None, min_samples_split = 2	5-fold CV F1; retained unpruned as a single-split baseline
LSTM	embedding dim $\in \{64, 128\}$; units $\in \{64, 128\}$; dropout $\in \{0.2, 0.5\}$; batch $\in \{32, 64\}$	Embedding 128 (trained from scratch), 1 LSTM layer of 128 units, dropout 0.2, dense sigmoid output, Adam (lr = 0.001), batch 64, 5 epochs with early stopping on validation loss (patience 2)	Validation loss on a 10% split of the training partition

Table 3. Dataset Composition and Split

Item	Value
Fake articles (Fake.csv)	23,481
Real articles (True.csv)	21,417
Rows kept after cleaning	44,856
Class balance	52.3% fake / 47.7% real
Training set	35,884 articles (80%)
Test set	8,972 articles (20%)
Label convention	1 = FAKE (positive class), 0 = REAL

3.5. Data Preprocessing

Because data preparation enhances the quality and usability of raw data, it is crucial for text analysis. It makes the text clearer and more meaningful by eliminating extraneous words, punctuation, and special characters. Preprocessing improves classification performance and increases the accuracy of machine learning models by standardizing and simplifying the data. Additionally, it prevents overfitting and boosts efficiency by reducing dimensionality through the removal of superfluous words and text normalization. Additionally, superior feature extraction techniques like TF-IDF and word embedding are supported by clean and organized data. The dataset was preprocessed in a number of ways to enhance text quality and get it ready for analysis. First, punctuation markers like commas, periods, exclamation points, and question marks were eliminated because these special characters typically add little significance to text classification and could produce

extraneous noise. To maintain consistency and stop the model from considering words like “News” and “news” as distinct tokens, all text was then transformed to lowercase. Additionally, frequently used stop words like “the,” “is,” “and,” and “in” were eliminated because the words have little semantic value and could increase the dimensionality of the data, which would decrease the effectiveness of the model. The text was then divided into smaller parts, typically words or tokens, using tokenization, which enabled the model to examine the data more thoroughly.

Additionally, words were reduced to their root or base forms using stemming and lemmatization. For example, “running,” “runs,” and “ran” were reduced to “run,” which helped standardize the text and enhance the detection of comparable phrases. The cleaned data were then converted into numerical representations using feature extraction approaches for model training and prediction. The pipeline

was applied in a fixed order: removal of URLs, HTML residue and non-alphabetic characters; case folding; tokenisation with NLTK's word tokeniser; removal of English stop words from the NLTK list; and lemmatisation with the WordNet lemmatiser, with stemming used only in the preliminary comparison and not in the reported runs, since lemmatisation preserved token readability required for SHAP interpretation. Records with empty or duplicate text bodies were dropped before splitting.

Vectorisation used scikit-learn's TfidfVectorizer restricted to the 5,000 highest-scoring unigrams, with sublinear term-frequency scaling, minimum document frequency of 5 and maximum document frequency of 0.7. The vectoriser was fitted on the training partition only and then applied to the test partition, so that no test-set vocabulary or document-frequency information entered the fitted representation. For the LSTM, texts were converted to integer sequences over a 10,000-token vocabulary and padded or truncated to the first 150 tokens.

3.6. Tools, Technologies and Programming Languages Used

A variety of programming languages, libraries, and development environments that facilitate ML and NLP activities are used in the creation of the FND system. These technologies assist in model training, model evaluation, feature extraction and data preprocessing, and all of the above were performed using Python. It provides various libraries to deal with datasets, develop models, and creating text analysis systems. Pandas is used for data manipulation, NumPy is used for numerical operations and matrix manipulation, Matplotlib/Seaborn is used for plotting data patterns and assessing model performance, Natural Language Toolkit (NLTK) is used for text data preprocessing (such as tokenization and stop-word removal), and Scikit-learn is used for implementing classification algorithms and evaluating their performance.

3.7. System Interface

The FND system includes a simple User Interface (UI) designed to allow users to input news text and receive immediate classification results. While the system's core functionality runs in the backend using Python and machine learning models, a lightweight interface was developed to improve usability and demonstrate practical application.

3.7.1. Index Page

The Index page contains the research title, student/research information and supervisor name.



Fig. 2 Index Page

3.7.2. Input Interface

The input interface allows the user to provide a news article or headline.

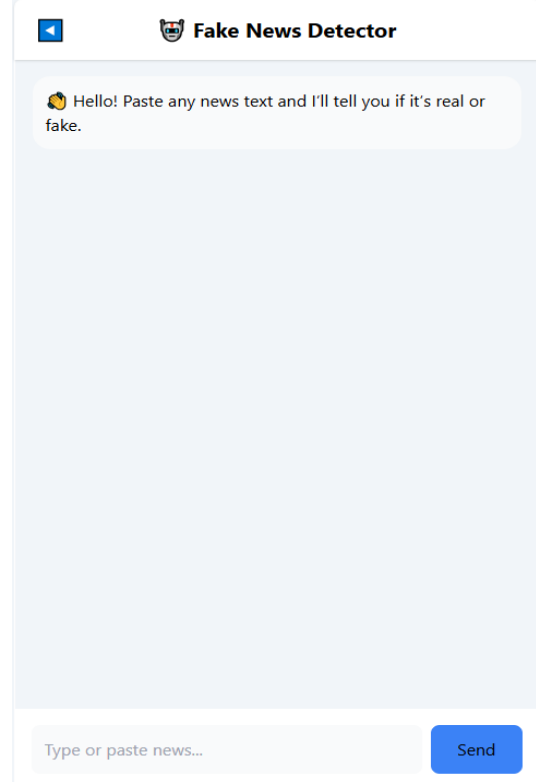


Fig. 3 Input Interface

3.7.3. Output Interface

Once the text is processed, the system displays the classification result on the output interface.

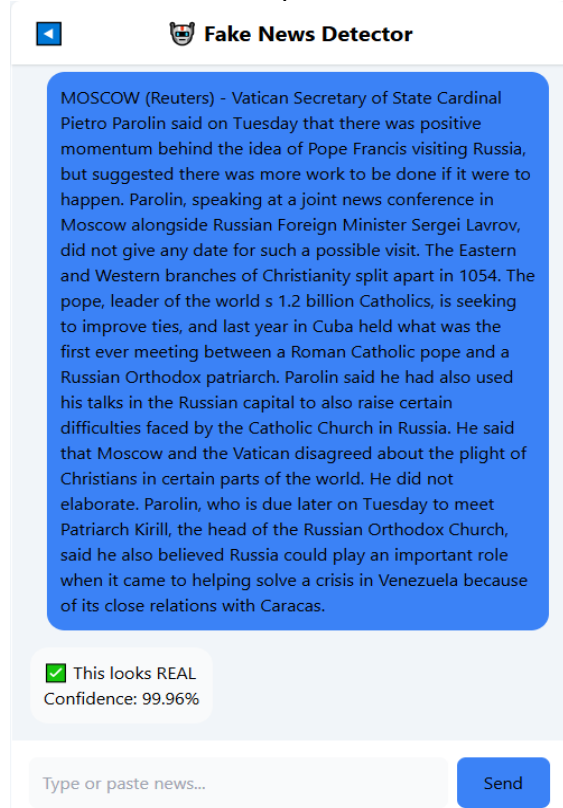


Fig. 4 Output Interface

4. Development of the Model

Algorithms used to build the model include Long Short-Term Memory (LSTM), SVM, and LR. The preprocessed dataset was divided into two parts; 80% was used for the training set, while the remaining 20% were used for the testing set. Models were evaluated with metrics such as accuracy, precision, recall and F1-score.

4.1. Model Results Presentation

The FND system's performance results are shown in this section. A testing dataset was used to assess how well the trained algorithms classified news stories as authentic or fraudulent. Each model was thoroughly evaluated using performance criteria like accuracy, precision, recall, and F1-score.

Table 4. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Logistic Regression	0.9905	0.9917	0.9902	0.9909	0.9992
Support Vector Machine	0.9921	0.9925	0.9923	0.9924	0.9996
Random Forest	0.9852	0.9886	0.9829	0.9858	0.9991
Decision Tree	0.9448	0.9589	0.9345	0.9465	0.9597
LSTM	0.9462	0.9334	0.9659	0.9494	0.9752

The experimental results, compared with each other, provide a comprehensive evaluation of the five predictive models for fake news detection. The models comprise one Deep Learning (DL) model, LSTM and four traditional Machine Learning (ML) algorithms, LR, SVM, RF and DT. The analysis of the results indicates that each model has acceptable predictive performance, which means that the use of supervised learning and Natural Language Processing (NLP) based feature engineering techniques is applicable in detecting misinformation automatically. Nonetheless, distinct variations were noted in the model's overall classification efficacy, discriminative power, and resilience. The win is the two linear models. TF-IDF already generates 5,000 features to represent each document, and the fake-versus-real separation is almost linear in this space, so the problem is best suited for linear separation. The single Decision Tree is the weakest tree model because it is required to choose one word for each split, and there is no single word that could determine

the class. The Random Forest overcomes this by averaging 150 different trees and gets to 98.5%. There are three practical reasons why the LSTM is not as high as the linear models, which are as follows: First, the LSTM only reads the first 150 words for each article; Second, the LSTM's embeddings are trained from scratch on approximately 36,000 documents, whereas the linear models' embeddings are pre-trained; Third, the LSTM was trained on a single CPU core, which limits the size of the model. It has a different error profile, but the highest recall (0.9659) and the most false positives. It aggressively blocks fake articles but is less inclined to mislabel a real article. That is not necessarily a bad deal if it is the more expensive error to make if you miss a fake article.

4.2. Confusion Matrix

The confusion matrix for the models illustrates how many news articles were correctly or incorrectly classified:

Table 5. Confusion Matrix

Model	True Neg	False Pos	False Neg	True Pos
Logistic Regression	4,245	39	46	4,642
Support Vector Machine	4,249	35	36	4,652
Random Forest	4,231	53	80	4,608
Decision Tree	4,096	188	307	4,381
LSTM	3,961	323	160	4,528

4.3. Interpretation of the Models Using SHAP

SHAP assigns every word in a document a number, and that number says how much that word moved the model's output for that specific document.

Positive SHAP value pushed the prediction toward FAKE, while negative SHAP value pushed the prediction toward REAL, and near-zero means the word made no difference. The values are additive: the base rate plus the sum of all word contributions equals the model's output for that document.

In plain terms, how much the word shifts the decision, averaged over all the articles, regardless of the direction it shifts, is the mean absolute SHAP value for that word, as seen in Figure 5. Words are prioritized in order of influence, with the most influential words at the top.

The colour of the bar indicates the direction (i.e., only on the articles in which the word is present, as highlighted below):

- red = when this word is encountered, it biases the prediction to FAKE
- blue = when this word is used, it pushes the prediction towards REAL.

Four of the five models contain 'said' as their most important word, and it is blue throughout. This is the most evident result in all the analysis. Professional news writing attributes statements constantly, the minister said, the report said. Fake articles make statements outright. In table four, the SHAP value of each word in the feature ranking was tabulated. The value of computing SHAP for every model, rather than for the winner only, lies in what the comparison

reveals. Agreement between architecturally dissimilar models on the same small set of high-attribution tokens indicates that the signal is a property of the corpus rather than an artefact of one inductive bias, which raises confidence that the learned rule is a genuine stylistic regularity. Where the models diverge, the divergence is diagnostic: the tokens with the largest attributions are stylistic and structural (attribution verbs, dateline city names, weekday names, and the media-embed markers “video”, “featured” and “image”) rather than propositional. Two consequences follow for deployment.

First, the classifier is partly a detector of publication format, so a fabricated article written in disciplined wire-service style is the predictable failure mode, and this should be stated to any operational user. Second, because named political entities such as “obama” and “hillary” carry non-trivial attribution towards FAKE, the model inherits the topical composition of the corpus, and periodic re-training with attribution monitoring is required if the system is not to encode the political period of its training data. Neither observation is recoverable from the accuracy table alone.

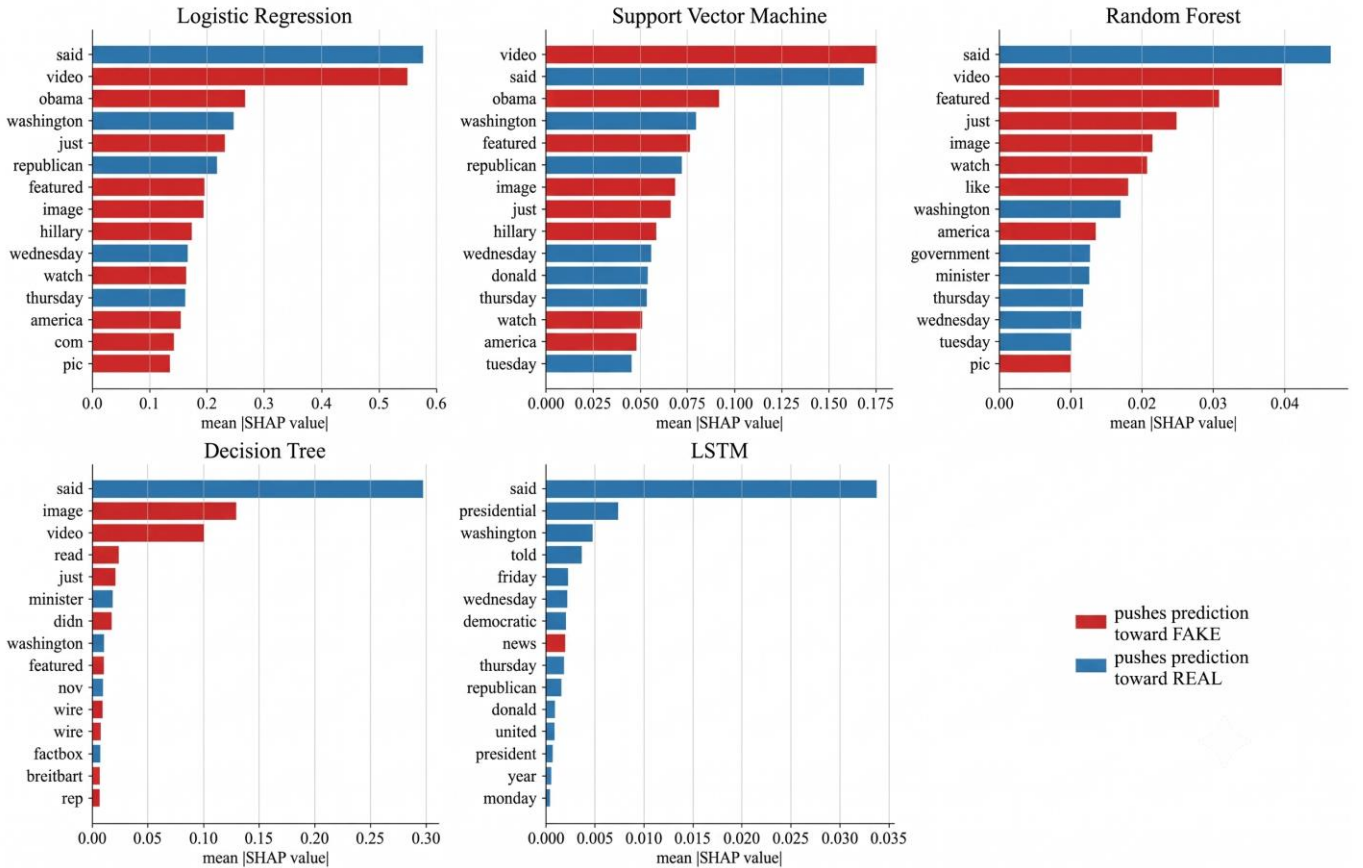


Fig. 5 SHAP feature importance ranking

Table 6. Words by SHAP Importance

Word	mean SHAP	Direction	Appears in
said	0.5773	REAL	73%
video	0.5485	FAKE	22%
obama	0.2663	FAKE	21%
washington	0.2459	REAL	27%
just	0.2297	FAKE	33%
republican	0.2165	REAL	27%
featured	0.1946	FAKE	21%
image	0.1939	FAKE	24%
hillary	0.1721	FAKE	16%
wednesday	0.1655	REAL	15%

4.4. Contributions of the Study

This study makes the following contributions. (a) It provides a model-agnostic explanation layer applied uniformly across five heterogeneous classifiers on one corpus, rather than explaining a single best model, which permits a cross-model comparison of decision evidence that has not

previously been reported for this dataset. (b) It reports a cross-model convergence finding: four of the five models rank the attribution token “said” as their most influential feature, and in each case the token pushes the prediction towards REAL, indicating that the classifiers are keying on attribution practice, a stylistic hallmark of professional reporting rather

than on topic alone. (c) It demonstrates that on a large TF-IDF representation, linear models (SVM 99.21%, LR 99.05%) match or exceed a from-scratch LSTM (94.62%), and it explains this outcome in terms of the near-linear separability of the feature space and the input truncation and embedding constraints of the LSTM, thereby providing an interpretable account of a result that is often reported without explanation. (d) It supplies a fully specified, reproducible protocol (class balance, split, vectoriser settings, hyperparameters) to support replication. (e) It delivers a working end-to-end prototype interface, linking the benchmark to a deployable artefact for newsroom and fact-checking use.

5. Conclusion

This research will be a significant contribution to fake news detection because it creates an integrated model of Natural Language Processing (NLP), Machine Learning (ML) and Explainable Artificial Intelligence (XAI) for accurate classification and interpretation of news articles as fake or real. Five algorithms (LR, SVM, RF, DT and LSTM) were compared on a common benchmark corpus under an identical preprocessing and partitioning protocol, with SVM attaining the strongest overall performance (99.21% accuracy; 0.9996 AUC-ROC). Attribution and formatting tokens were then calculated for all five models, which resulted in a converging decision basis for the classifiers, a basis that is linguistically interpretable, and identified the main weakness of the framework: fabricated content that mimics professional reportorial style. Among the practical implementations of the

proposed framework to journalists, fact checkers, media and digital platforms, one of the suggestions is the development of an easy-to-use system that detects fake news. The study shows that the integration of XAI and ML into FAD systems can lead to more reliable, transparent, and applicable systems. The findings show that robustly engineered features and preprocessing, along with both traditional machine learning and deep learning methods, can provide reliable and scalable approaches to automated misinformation detection. Given this situation, the study concludes that intelligent FND systems could be of enormous importance in assisting digital platforms, media and the legislators to detect misinformation, earn public trust and improve evidence-based decision making. While these promising findings underscore the value of leveraging deep learning in NLP, future research should delve deeper into these transformer architectures, such as BERT, RoBERTa, and GPT-based models, to explore the more intricate contextual semantics that can be conveyed. More studies on multilingual and cross-domain data sets are desired to promote generalization across languages and reduce misinformation narratives. Moreover, the use of multimodal data, including photos, videos, trustworthiness signals, and user engagement patterns, may further enhance the accuracy of the detection system. Enhancing accountability, transparency, and user confidence all depend on the adoption of Explainable AI approaches. Lastly, to guarantee long-term efficacy against quickly changing disinformation tactics, real-time deployment and adaptive continuous-learning systems should be investigated.

References

- [1] Hadeer Ahmed, Issa Traore, and Sherif Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 127-138, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Nipa Akter et al., "Advanced Detection and Forecasting of Fake News on Social Media Platforms using Natural Language Processing and Artificial Intelligence," *Journal of Posthumanism*, vol. 5, no. 6, pp. 3208-3236, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo, "A Comparative Study of Machine Learning and Deep Learning a Techniques for Fake News Detection," *Information*, vol. 13, no. 12, pp. 1-28, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Hunt Allcott, and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211-236, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Emad Alsuwat, and Hatim Alsuwat, "An Improved Multi-Modal Framework for Fake News Detection using NLP and Bi-LSTM," *The Journal of Supercomputing*, vol. 81, no. 1, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Isabel Bezzaoui et al., "Explainable AI for Online Disinformation Detection: Insights from a Design Science Research Project," *Electronic Markets*, vol. 35, no. 1, pp. 1-28, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, 2009. [[Google Scholar](#)]
- [8] Jawaher Alghamdi, Suhuai Luo, and Yuqing Lin, "A Comprehensive Survey on Machine Learning Approaches for Fake News Detection," *Multimedia Tools and Applications*, vol. 83, pp. 51009-51067, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse, "A Survey of Explainable AI Techniques for Detection of Fake News and Hate Speech on Social Media Platforms," *Journal of Computational Social Science*, vol. 7, no. 1, pp. 587-623, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Haewoon Kwak et al., "What is Twitter, a Social Network or a News Media?," *Proceedings of the 19th International Conference on World Wide Web*, Daejeon, Korea, pp. 591-600, 2010. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] David M.J. Lazer et al., "The Science of Fake News," *Science*, vol. 359, no. 6380, pp. 1094-1096, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Pooja Malhotra, and S.K. Malik, "Fake News Detection using Ensemble Techniques," *Multimedia Tools and Applications*, vol. 83, no. 14, pp. 42037-42062, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Silvi Munawaroh, "Explainable AI (XAI) for Fake News Detection: A Review of Interpretability in Deep Learning Models for Misinformation Classification," *International Journal of Research and Applied Technology*, vol. 4, no. 2, pp. 334-340, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [14] K. Neela, K. Deekshitha, and R. Deepa, "An Ensemble Learning, Frame Work for Robust Fake News Detection," *International Journal of Engineering Research and Technology*, vol. 13, no. 1, pp. 1-6, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Halyna Padalko et al., "Ensemble Machine Learning Approaches for Fake News Classification," *Radioelectronic and Computer Systems*, no. 4, pp. 5-19, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Badugu Akshitha Priya et al., "Fake News Detection using Natural Language Processing," *International Journal of Computer Technology and Electronics Communication*, vol. 9, no. 2, pp. 541-550, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Hannah Rashkin et al., "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2931-2937, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Saurabh Pratap Singh Rathore et al., "Comparative Analysis of Machine Learning Models for Fake News Detection using a Natural Language Processing," *2024 1st International Conference on Advances in Computing, Communication and Networking*, Greater Noida, India, pp. 1389-1393, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Kai Shu et al., "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22-36, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] William Yang Wang, "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 422-426, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo, "Cross-Domain Fake News Detection using a Prompt-Based Approach," *Future Internet*, vol. 16, no. 8, pp. 1-16, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Xinyi Zhou, and Reza Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1-40, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]