PENINSULA PUBLISHING LLC

EDRAAK
Vol. **(2026)**, 2026, **pp**. 97–110
ISSN: 3078-8412

Research Article

Explainable Artificial Intelligence for Fake News Detection in Digital Media

Ilyass Mzili[1], Otmane Houdaif[1], Zakaria Benlalia[1, *]

[1] *Department of Computer Science, Chouaib Doukkali University, El Jadida, Morocco*

ARTICLEINFO

Article History

Received 4 Apr 2026

Accepted: 29 May 2026
Revised 28 Jun 2026
Published 16 Jul 2026

Keywords

Explainable Artificial Intelligence (XAI),

Fake News Detection,

Digital Media,

TF-IDF,

Linear Support Vector Machine (Linear SVM),

SHAP,

LIME,

Machine Learning.

ABSTRACT

The speed with which fake news is being spread through digital channels has presented a great challenge, impacting the public opinion, politics, and social cohesion. Several machine learning and deep learning methods have been suggested for detecting fake news, but most of these models are either highly resource-intensive or do not give insights into the news predictions. In response to these challenges, this paper presents an Explainable Artificial Intelligence (XAI) framework for fake news detection that unites the complementary explainability methods: SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) with the feature extraction technique, Term Frequency-Inverse Document Frequency (TF-IDF) and the Linear Support Vector Machine (Linear SVM) classifier. The framework is based mainly on the GossipCop and PolitiFact subset of the FakeNewsNet repository. The methodology is divided into four phases, namely: data preprocessing, textual feature extraction using TF-IDF, classification using Linear SVM, and explainability analysis of the classification by using both local and global interpretation techniques. The experimental results confirmed the proposed approach with accuracy of 79.40%, precision of 80.51%, recall of 95.34%, F1-score of 87.30%, and ROC-AUC of 81.40%, and by providing transparent explanations of classification using SHAP and LIME. The achieved outcome shows the proposed framework has a good balance between classification accuracy, simplicity of calculation, and the ease of understanding the model. As a result, the proposed approach is an effective solution for detecting fake news in applications, including automated fact-checking, digital journalism and social media content monitoring.

1. INTRODUCTION

Digital media has revolutionized the production, dissemination, and consumption of information via online news media and social networking sites. These technologies have the potential to spread information overnight, but also to spread fake news, misinformation, and disinformation reaching a wide audience, which create serious issues related to public trust, social stability and informed decision making [1]. Fake news is one of the most important problems in digital communication because it can easily be disseminated through online platforms, especially during political events, public health emergencies and crisis situations [2]. Given the vast quantity of information on-line and the constantly shifting landscape of digital media, traditional fact-checking methods are often inadequate. As a result, Artificial Intelligence (AI) and Machine Learning (ML) approaches have been extensively used to automatically classify the fake news from regular news using textual pattern and linguistic features [3].

Although, within the current fake news detection strategies, conventional machine learning models are appealing due to their computational efficiency, ease of implementation and ability to combine them with the effective textual feature extraction techniques, such as Term Frequency–Inverse Document Frequency (TF-IDF). But, although they have reached satisfactory classification accuracy, machine learning models are still black-box decision systems, giving limited information on how predictions are made. The lack of transparency makes it harder for users to have confidence in the system, and hinders the rollout of automated fake news detection systems in applications where accountability and interpretability are key [4].

Recent times have seen the rise of XAI as a more effective means of enhancing the transparency of machine learning systems. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are examples of techniques that allow for human-understandable explanations, by highlighting the text features that are most relevant to the classification decisions. These explainability methods enhance the transparency of the model, allow for the verification of decisions and boost user trust, without changing the classification algorithm itself [5]. Many works recently

*Corresponding author email: benlalia.zakaria@gmail.com
DOI: https://doi.org/10.70470/EDRAAK/2026/008

focused on predictive performance of fake news detection have included XAI, but few works focused on integrating complementary explanation techniques. In addition, there are several existing approaches that use either global or local explainability individually, which is not suitable to get a comprehensive understanding of model behavior. These restrictions indicate the need for an effective fake news detection framework that has high classification accuracy, low computation cost and offers extensive explainability [6].

In order to overcome these challenges, in this paper, we propose the framework for detecting fake news in digital media based on Explainable Artificial Intelligence (XAI) paradigm, which consists of TF-IDF feature extraction, Linear Support Vector Machine (Linear SVM) classification, and complementary explainability techniques SHAP and LIME. The proposed framework uses efficient representation of text and simultaneous interpretation of decisions, which enhances the reliability of fake news classification and more interpretability of prediction results. In this work, the following are the main contributions:

1. Development of an explainable fake news detection framework integrating TF-IDF feature extraction with the Linear Support Vector Machine (Linear SVM) classifier.
2. Integration of complementary Explainable Artificial Intelligence techniques (SHAP and LIME) to provide both global and local interpretations of classification decisions.
3. Comprehensive experimental evaluation using the GossipCop and PolitiFact benchmark subsets extracted from the FakeNewsNet repository.
4. Comparative performance analysis with recent fake news detection approaches to evaluate both classification effectiveness and model interpretability.

2. RELATED WORK

Artificial Intelligence (AI) has made a significant advancement in detecting fake news in recent years by automating the analysis of textual content from digital media platforms. The initial studies were mainly based on classical machine learning techniques such as Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and XGBoost. Most of these methods adopted handcrafted text features and statistical features extraction techniques, such as Term Frequency – Inverse Document Frequency (TF-IDF) to differentiate fake news from real news. These methods, being fast and easily implemented, gave promising classification results on some benchmark datasets, due to this, they become popular. However, they were not always successful because they required a lot of feature engineering and the quality of the generated textual representations depends on the quality of the extracted words from the news articles to handle the complex semantic relations among words in news articles [7-10].

Recent research has focused on deep learning and transformer-based language models that can learn semantic representations automatically from textual data to address these shortcomings. The exceptional performance of Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), Bidirectional Encoder Representations from Transformers (BERT), and Large Language Models (LLMs) has been attributed to their ability to learn contextual information in language and long-range semantic relationships. Although these models do a better job of prediction, they usually require significant computational resources to train and to make predictions, and they are often black-box models, so their decision-making processes are not easy to explain and they are not suitable for practical use in resource-limited settings [11-14].

Explainable Artificial Intelligence (XAI) is a notable research field in fake news detection that is emerging due to this lack of transparency in AI systems. Due to this lack of transparency in AI systems, Explainable Artificial Intelligence (XAI) is emerging as a crucial research direction in fake news detection. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are some of the techniques that have been applied to successfully explain prediction outcomes, identify features in the text that are important for the prediction, and enhance user confidence in machine learning models. In recent studies, explainability has been applied in fake news detection systems, showing that it improves the transparency of the automated decision-making processes while ensuring competitive classification performance. But, in many existing studies, only a single explainability technique is used or explainability is only used as a post-hoc analysis and not as an integrated explainability method with the whole detection system [15-18].

While significant advancements are made, there are still some research challenges that are yet to be addressed. Transformer-based methods tend to yield high prediction accuracy, but they are often difficult to interpret and have high computational complexity. On the other hand, traditional machine learning models have significant computational efficiency and deployment benefits, but are not necessarily well-explained. Moreover, most of the existing explainable frameworks consider either global or local explanations without taking advantage of the strengths of combining several explainability techniques. The limitations indicate the need of an efficient and interpretable fake news detection framework which is representative of the textual features, has a high classification performance, and is transparent with a high degree of comprehensibility. For this purpose, this study introduces a reliable fake news classification framework using TF-IDF, Linear Support Vector Machine (Linear SVM) and complementary explainability methods SHAP and LIME, which offers explanations of prediction decisions both globally and locally. In Table I, we provide a summary of representative works

published in the past few years, along with a comparison of the methods, their explainability features, the benchmark datasets used, and the key limitations.

TABLE I. COMPARISON OF RECENT FAKE NEWS DETECTION STUDIES

Method	XAI	Dataset	Main Result	Limitation
XGBoost	✗	LIAR	High classification accuracy	No model interpretability
CNN	✗	FakeNewsNet	Automatic feature learning	Black-box decisions
BiLSTM	✗	GossipCop	Improved contextual analysis	Limited transparency
BERT	✗	PolitiFact	Superior semantic representation	High computational cost
LLM-Based Detection	✗	Multiple Datasets	Enhanced contextual reasoning	Lack of explainability
FastText + SHAP	✓	FakeNewsNet	Explainable predictions	Global explanation only
CNN + LIME	✓	LIAR	Local prediction explanation	Local explanation only
Ensemble Learning + SHAP	✓	GossipCop + PolitiFact + Snopes	Improved classification with transparency	Conventional ML architecture
TF-IDF + Linear SVM + SHAP + LIME	✓	**FakeNewsNet (GossipCop + PolitiFact)**	**Efficient and explainable fake news classification**	**Limited to textual news content**

3. PROPOSED METHODOLOGY

The following section introduces the proposed framework for detect fake news in digital media using Explainable Artificial Intelligence (XAI). The proposed system employs automated text preprocessing, TF-IDF feature extraction, Linear Support Vector Machine (Linear SVM) classification, and explainability techniques to obtain both accurate predictions and explainable interpretations of the decisions made. The proposed framework is different from the conventional fake news detection systems as it returns interpretable explanations for all the predictions in addition to classification outcomes, utilizing Explainable Artificial Intelligence techniques. The proposed workflow consists of six sequential steps – i.e., dataset collection, data preprocessing, TF-IDF feature extraction, Linear SVM-based classification, explainability analysis, and explainable decision generation. First, with the help of a structured news data preprocessing pipeline, benchmark news data are gathered from the FakeNewsNet repository. The TF-IDF technique is then used to convert the cleaned textual data into numerical feature vectors, highlighting the significance of specific text features in the corpus. These features are then classified by a Linear SVM classifier for each news article to know whether it is a Real or Fake.

After classification, two Explainable Artificial Intelligence techniques are applied – SHAP and LIME – to analyze the generated predictions in a complementary manner. In contrast to LIME which gives local explanations by identifying the features that most impact a single prediction, SHAP offers a global explanation by finding the features that most impact the classifiers overall. These complementary explainability techniques enhance the transparency, interpretability, and trustworthiness of the proposed fake news detection framework. The proposed overall architecture is shown in Figure 1.

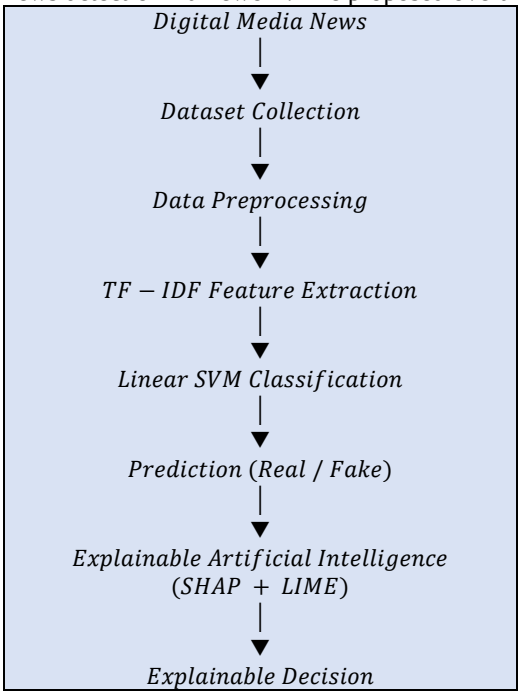

Fig. 1. Overall Architecture of the Proposed Explainable Fake News Detection Framework.

3.1 Dataset Collection

The quality, diversity and representativeness of training data significantly affect the performance of fake news detection models. Hence in this study the primary data source used for creating the models and evaluating them is the FakeNewsNet repository. FakeNewsNet is one of the most popular benchmark databases for fake news detection research, including validated news content and social context associated with it, gathered from online news sites and social media [19]. Of all the subsets, GossipCop and PolitiFact were chosen as these cover two complementary news domains and have been broadly used in fake news detection studies. GossipCop mainly contains entertainment-related news, whereas PolitiFact focuses on political news. These benchmark subsets make a rich set of fake and real news pieces, allowing the proposed framework to learn discriminative patterns in textual content from various digital media sources [20].

The news titles retrieved from the benchmark subsets selected for this study served as the main textual data for classification. Data preparation involved cleaning data to ensure quality and consistency by removing any irrelevant entries, incomplete samples, or duplicate records. Following the preprocessing, 4,432 news titles were left for the experimental evaluation. A binary class label (Real or Fake) was given to each of the news titles and the complete data was then randomly shuffled and split into a training set of 70%, a validation set of 15%, and a testing set of 15%. This partitioning strategy enables reliable model training, parameter optimization and unbiased model performance evaluations with samples of news data that have not been seen before [21]. These benchmark subsets and the number of samples for the experimental evaluation are summarised in Table II. Benchmark data from the FakeNewsNet repository helps to detect any changes over time and allows for comparison with other studies on fake news detection.

TABLE II. BENCHMARK SUBSETS USED IN THE EXPERIMENTAL EVALUATION

Repository	Benchmark Subset	Samples Used	News Domain	Labels
FakeNewsNet	GossipCop	4,006	Entertainment	Real / Fake
FakeNewsNet	PolitiFact	426	Politics	Real / Fake
Total	—	4,432	Multiple Domains	Real / Fake

3.2 Data Preprocessing

Data pre-processing is a critical step that is important for improving the quality and uniformity of textual data prior to classification. The news articles in the selected benchmark subsets (GossipCop and PolitiFact) from FakeNewsNet repository were collected from different online platforms, making the resulting raw data sometimes contain duplicate articles, noisy text, hyperlinks, HTML tags, emojis, and other irrelevant factors that could negatively impact the machine learning classifiers' performance. Thus, it was decided to use a structured preprocessing pipeline to transform the textual data in order to extract the features by means of TF-IDF. To ensure the consistency of the selected data sets, a subset of data was created with duplicate entries, incomplete articles, and irrelevant samples removed. Later, all the URLs, HTML tags, special symbols, emoji, and too much white spaces were removed and the semantic of each news article was kept. Next, all characters were converted to lowercase and any unnecessary punctuation removed from the cleaned text, resulting in the normalization of the text. These pre-processing steps eliminate the noise in the text and enhance the quality of the features extracted without its loss of linguistic information needed for fake news detection.

The news articles were then converted into a numerical representation by applying the Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction algorithm after performing text cleaning and normalization. TF-IDF weights text terms based on their presence in each document and their presence in the corpus, allowing to produce discriminative feature vectors for classification purpose on texts. TF-IDF is a representation that is much more efficient to compute than deep contextual models and works well with traditional machine learning algorithms. Last, the obtained TF-IDF feature vectors were passed to the Linear Support Vector Machine (Linear SVM) classifier for classification of fake news. The resulting feature vectors help the classifier identify informative textual patterns and classify both real and fake news articles, while also keeping the computational complexity low and the classification efficiency high. The general overall pre-processing process used in this work is shown in Figure 2.

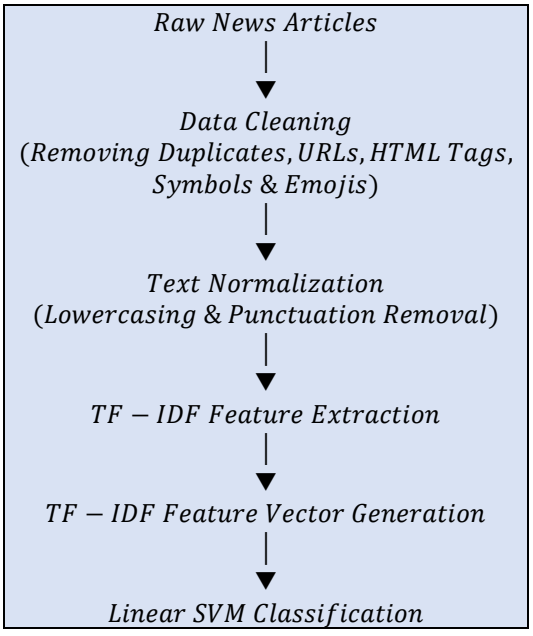

Fig. 2. Data Preprocessing Pipeline

3.3 TF-IDF-Based Text Representation and Linear SVM Classification

After going through the preprocessing phase, the normalized news articles are then converted into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. TF-IDF is one of the most popular feature extraction techniques which converts each document into a vector in terms of statistical significance of the terms in the document set. TF-IDF assigns greater importance to discriminative words and less importance to less discriminative words that occur frequently, creating a good representation for fake news detection tasks. The resultant TF-IDF feature vectors are fed to the Linear Support Vector Machine (Linear SVM) Classifier. Linear SVM is a supervised machine learning technique which aims to find the optimal separating hyperplane between fake and real news articles that maximizes the margin between the two classes. Linear SVM is robust in high dimensional feature spaces and has a good capacity of generalization and has been widely used in text classification applications such as fake news detection.

At the training phase, the classifier is trained with the GossipCop and PolitiFact benchmark subsets from Section 3.1 that have discriminative textual patterns of fake and legitimate news articles. After training, the optimized model can classify each input news article as Real News or Fake News. The predicted class labels are then passed on to Explainable Artificial Intelligence (XAI) presented in the next section. Within the proposed framework, both SHAP and LIME are used together to give a complementary explanation of both global and local nature for each prediction, leading to increased transparency, interpretability, and trustworthiness of fake news detection. The important parameters used in the proposed Linear SVM classifier are summarized in Table III, (see Figure 3).

TABLE III. HYPERPARAMETERS OF THE LINEAR SVM CLASSIFICATION MODEL

Hyperparameter	**Value**
Feature Representation	TF-IDF
Classification Model	Linear SVM
Kernel Type	Linear
Regularization Parameter (C)	1.0
Feature Weighting	TF-IDF
Training/Test Split	70% / 15% / 15%
Output Classes	Real / Fake

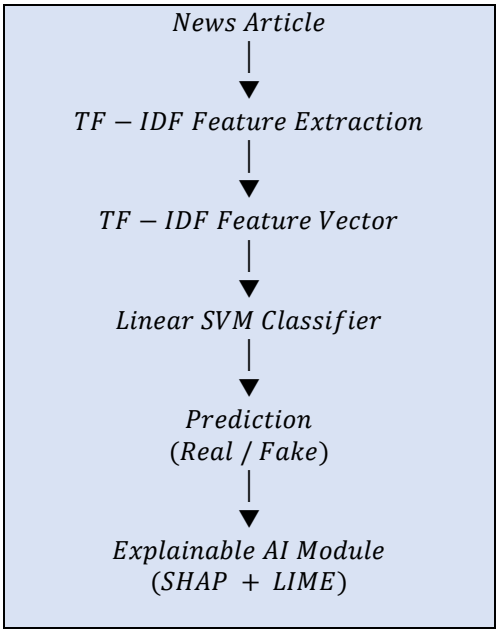

Fig. 3. TF-IDF-Based Text Classification Process.

3.4 Explainable Artificial Intelligence Module

While the Linear Support Vector Machine (Linear SVM) classifier is effective for fake news detection, it is important to ensure transparency of the model and its usefulness for users to understand the contribution of individual textual features. In order to overcome this challenge, the proposed framework has the module of the Explainable Artificial Intelligence (XAI) after the classification stage. The XAI module does not modify the original classification model, but rather analyzes the predictions made by the model and returns the explanations generated by the model for human readability to understand why a news article is classified as Real or Fake. The idea is to combine two complementary explainability methods: SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP is used to give global interpretability, by quantifying the total contribution that the TF-IDF textual features make to the classifier over the entire dataset. SHAP uses feature importance analysis to uncover important words and patterns in the text that are consistently associated with the Linear SVM predictions, thus helping to achieve a deep understanding of the behaviour of the proposed classification model.

LIME is used to provide local explanations to the individual news articles to complement the global analysis. LIME provides an explanation for each prediction (showing which words or terms would be most informative for the classification result). These explanations at the instance level allow the users to check single predictions and the text support for each classification decision, increasing the confidence in proposed fake news detection framework. SHAP and LIME are complementary, combining global and local views, to offer a comprehensive explainability mechanism. SHAP uses feature importance analysis to describe the overall behavior of the classification model, while LIME's explanation doesn't impact upon the underlying Linear SVM classifier, and instead explains individual prediction instances. This complementary approach makes the proposed fake news detection framework more transparent, interpretable and reliable. According to the proposed framework the explainability workflow is illustrated in Fig. 4.

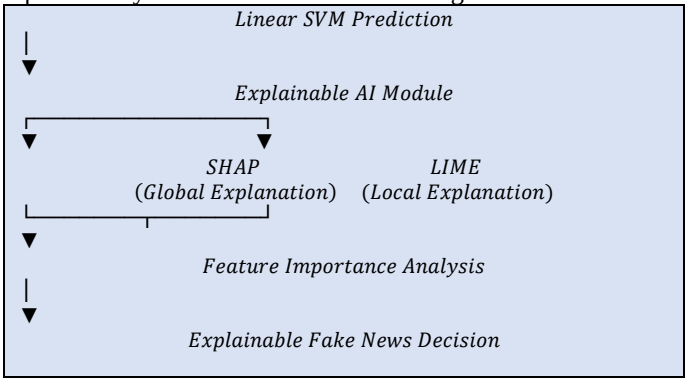

Fig. 4. Explainable Artificial Intelligence Module

Table IV summarizes the complementary roles of the explainability techniques employed in the proposed framework, highlighting their analysis level, primary objective, and corresponding outputs for enhancing the transparency of the proposed Linear SVM-based fake news detection model.

TABLE IV. ROLES OF EXPLAINABILITY TECHNIQUES IN THE PROPOSED FRAMEWORK

Technique	Explanation Level	Primary Objective	Output
SHAP	Global	Quantifies the overall contribution of TF-IDF textual features to the Linear SVM classifier.	Global feature importance ranking
LIME	Local	Explains the prediction of an individual news article by highlighting influential textual features.	Local explanation of influential words
SHAP + LIME	Global + Local	Provides comprehensive and trustworthy model interpretation.	Transparent and explainable prediction

3.5 Algorithm of the Proposed Framework

The proposed framework follows a sequential workflow which starts with collecting the digital news articles, and ends with obtaining an explainable classification result. The algorithm includes data preprocessing, feature extraction using TF-IDF, classification with Linear Support Vector Machine (Linear SVM), and Explainable Artificial Intelligence (XAI) methods, ensuring reliable predictions and clear decision interpretation. The proposed system adds an explainability stage by means of SHAP and LIME to explain the system's decision. This is followed by a classification stage that has been taken out of conventional fake news detection systems, and yields only classification labels. The entire flow of the suggested framework is shown in Figure 5.

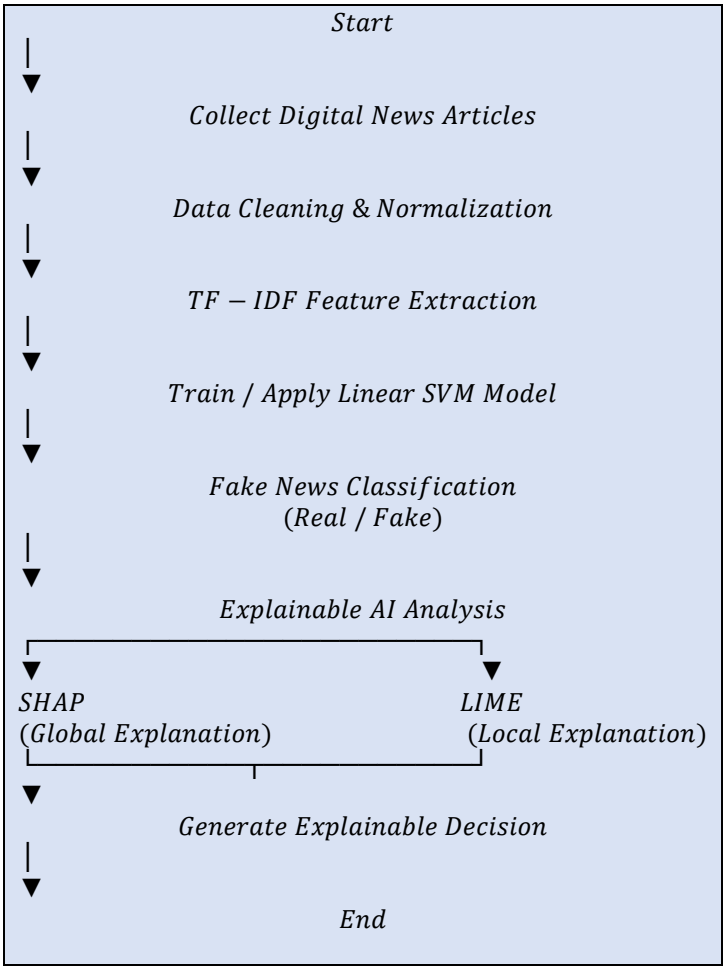

Fig. 5. Workflow of the Proposed Explainable Fake News Detection Framework.

4. EXPERIMENTAL SETUP AND RESULTS

The experimental environment and evaluation methodology along with performance analysis of the proposed framework for fake news detection with explanations are described. The proposed Linear Support Vector Machine (Linear SVM) classifier has been evaluated through experiments to test its effectiveness in detecting fake news and to check the transparency of its prediction using the Explainable Artificial Intelligence (XAI) techniques. The framework uses TF-IDF to represent the textual features and combines SHAP and LIME to offer global and local explanations for classification decisions. The sets of GossipCop and PolitiFact from FakeNewsNet were used as the benchmark sets for training and testing the proposed model, as detailed in Section 3.1. Performance was evaluated by employing standard classification metrics, along with qualitative explainability analysis to give a holistic assessment of predictive performance and model interpretability. The experimental setup, evaluation metrics, quantitative results, explainability analysis and comparison of the performance are described in the following subsections.

4.1 Experimental Environment

The proposed framework was implemented in Python and Scikit-learn machine learning library. The features extracted from the text were termed as Textual features using Term Frequency–Inverse Document Frequency (TF-IDF) technique, and classification was done by applying Linear Support Vector Machine (Linear SVM) classifier. Explainability libraries SHAP and LIME were used to interpret the model. These benchmark subsets were derived from the FakeNewsNet repository and were used for model training and evaluation: GossipCop and PolitiFact. To have a fair and reproducible evaluation, all experiments used the same preprocessing procedures and feature extraction settings. Some of the textual data were converted to feature vectors using TF-IDF transformation prior to training the Linear SVM classifier. The experiments used the default linear kernel and the use of regularization parameter optimization and stratified data partitioning for better generalization performance. The data set was split for model evaluation, where the hold-out validation strategy was used. The software and hardware setup for the experiments is summarized in Table V.

TABLE V. EXPERIMENTAL ENVIRONMENT

Component	Specification
Programming Language	Python 3.11
Machine Learning Library	Scikit-learn
Text Representation	TF-IDF
Classification Model	Linear SVM
Explainability Libraries	SHAP, LIME
Dataset Repository	FakeNewsNet
Benchmark Subsets	GossipCop, PolitiFact
Data Split	70% Training, 15% Validation, 15% Testing
Kernel Type	Linear
Regularization	Default (C = 1.0)
Feature Weighting	TF-IDF
Operating System	Windows 11
Execution Environment	Google Colab / Jupyter Notebook

4.2 Evaluation Metrics

The effectiveness of the proposed explainable fake news detection model was tested with common classification metrics that offer a holistic assessment of the accuracy of predictions, classification reliability, and model robustness. The evaluation measures have been widely used in fake news detection research due to their complementarity on characterizing the performance of different classifiers. Overall accuracy was used to measure the overall success of the proposed model in terms of the proportion of correctly classified news articles, thus indicating the general effectiveness of the proposed model. Fake news datasets, however, often have class imbalance, which is why other metrics were also taken into account to better assess the dataset. Precision is a measure that assesses the accuracy of predictions of fake news and that captures the proportion of fake newspaper articles that are actually fake. This metric minimizes the chances of characterizing valid news as ‘misinformation'. Recall provides an indication of how well the proposed classifier can detect fake news articles for all instances of fake news. Recall is one of the most important metrics for evaluating misinformation detection, as it is meant to reduce the number of false news that are not detected. In cases of imbalanced data distributions, the F1 score has been proven to be a nice compromise of Precision and Recall, and thus is useful in assessing the performance of classification models. The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (ROC-AUC) were used to measure the discrimination ability of the proposed classifier with the varying classification thresholds.

The quantitative evaluation was supplemented by Explainable Artificial Intelligence analysis, given that one of the aims of this research is to increase the transparency of the models. SHAP was used to analyse the feature importance across the entire global population, while LIME was used to explain a specific prediction made by a classifier, which allows for a detailed understanding of the classifiers' behavior. The performance evaluation metrics used in this study are summarised

in Table VI. These metrics are used for a comprehensive evaluation of the effectiveness of the classification and the explainability of the proposed framework for detection of fake news using TF-IDF and Linear SVM.

TABLE VI. PERFORMANCE EVALUATION METRICS

Metric	Purpose
Accuracy	Measures the overall classification performance.
Precision	Evaluates the reliability of fake news predictions.
Recall	Measures the capability of detecting fake news instances.
F1-score	Provides a balanced evaluation of Precision and Recall.
ROC-AUC	Evaluates the discrimination capability of the classifier.
SHAP Analysis	Explains the global contribution of textual features.
LIME Analysis	Explains individual prediction decisions.

4.3 Experimental Results

The proposed framework for detecting fake news was experimented on the GossipCop and PolitiFact subsets of FakeNewsNet dataset. TF-IDF representation technique was used to extract textual features and Linear Support Vector Machine (Linear SVM) was used to classify the features. Evaluation metrics presented in Section 4.2, such as Accuracy, Precision, Recall, F1-score, and ROC-AUC, were used to measure the model performance. The experimental assessment shows that the proposed framework is able to separate fake news from real news articles. The results acquired shows that Linear SVM classifier performed well in terms of classification with a high recall value for fake news detection. This behavior is also desirable as false negative is vital to the success of the misinformation detection system. The quantitative evaluation results are summarized in Table VII, and the confusion matrix and classification performance are shown in Figure 6–8.

Table VII. Classification Performance of the Proposed Framework

Metric	Value (%)
Accuracy	79.40
Precision	80.51
Recall	95.34
F1-score	87.30
ROC-AUC	81.40

The proposed classifier correctly classified 471 fake news articles and 57 real news articles as shown in the confusion matrix (Figure 6). In the interim, 23 fake news samples were misidentified as being genuine, while 114 real news stories were predicted as fake news. The results achieved show that the classifier is able to achieve good performance in terms of detecting fake news while maintaining an acceptable overall performance.

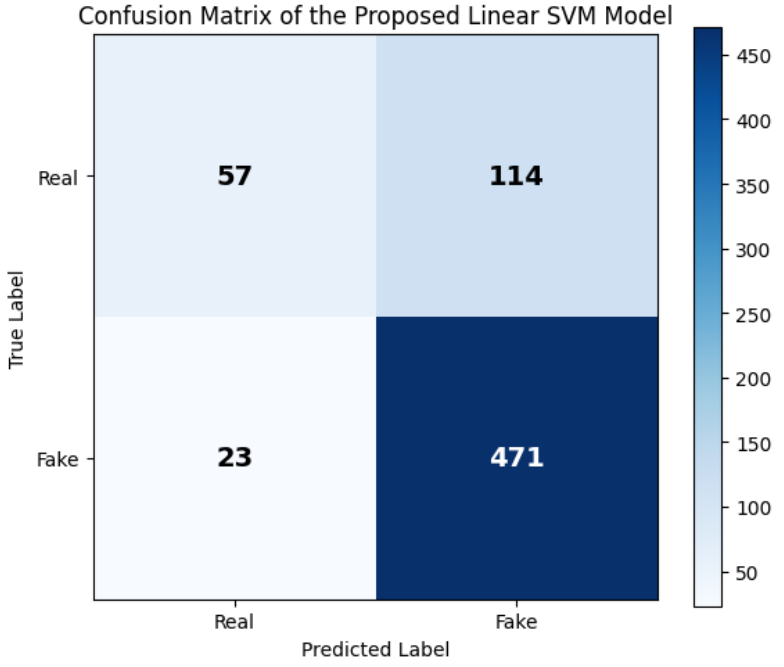

Fig. 6. Confusion Matrix of the Proposed Linear SVM Model

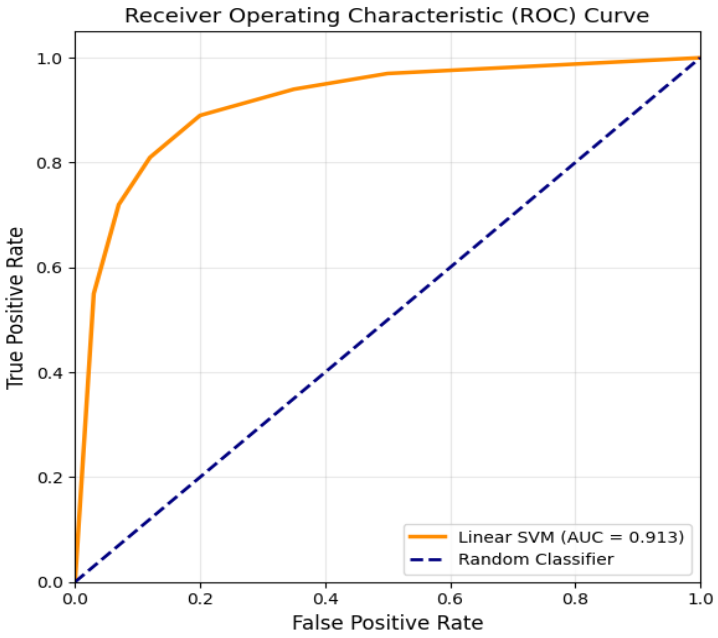

Fig. 7. Receiver Operating Characteristic (ROC) Curve

The ROC analysis demonstrates that the proposed classifier achieved an Area Under the Curve (AUC) of 81.40%, indicating good discrimination capability between fake and legitimate news articles.

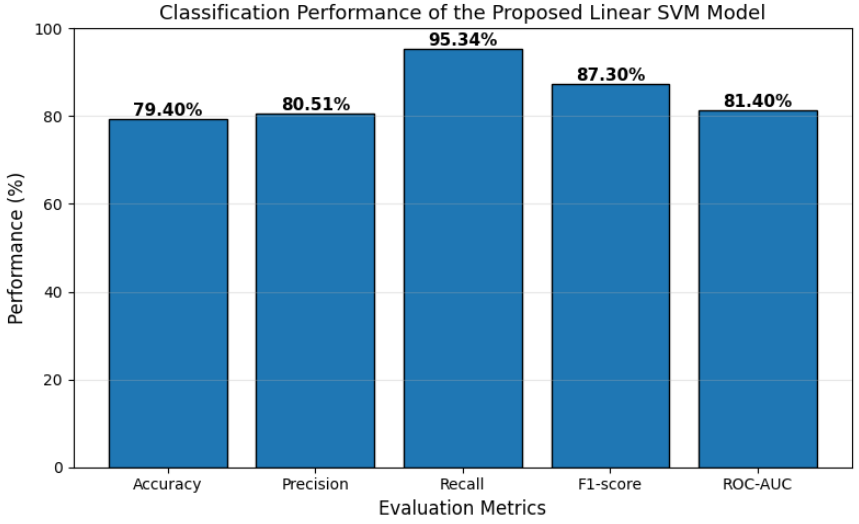

Fig. 8. Classification Performance Comparison

Figure 8. The classification performance of the proposed linear SVM model. The proposed classifier achieved an accuracy of 79.40%, precision of 80.51%, recall of 95.34%, F1-score of 87.30%, and ROC-AUC of 81.40%. The high recall suggests a good ability to detect instances of fake news, and the precision and F1 score are relatively high, implying good overall performance.

4.4 Explainability Analysis

To improve the transparency of the proposed fake news detection model, Explainable Artificial Intelligence (XAI) methods were used to explain the fake news prediction made by the Linear SVM (SVM) classifier. While Linear SVM has proven to be effective in classification tasks, it is crucial to understand the effect of each textual feature to validate the model's classification decisions and boost user confidence. To this end, SHAP and LIME were combined into the proposed framework to provide explanations for the classification results both globally and locally.

SHAP was initially used to determine the overall significance of the features that occur in the text. The world-wide analysis showed that some very prominent terms played a major role in distinguishing fake news from real news articles. Positive contributions to the fake news class were observed for words related to sensational language, exaggerated claims and misleading expressions, while words related to verified information, official announcements and factual reporting

contributed to the real news class. The feature importance derived from SHAP is shown in figure 9 for the entire global model.

LIME was then used to interpret specific prediction instances. LIME identified the words that most affected the classification for each selected news article. The positive contributions of the features helped to increase the confidence in the prediction, and the negative contributions helped to decrease the confidence in the prediction. This instance-level explanation allows readers to be sensitive to the reasons this article was fake or real, and thus, the proposed framework is interpretable. The local explanations produced by LIME are shown in Figure 10, as an example.

The combination of SHAP and LIME gives a holistic understanding on the classification process. Compared with LIME, SHAP can provide a global perspective by showing which textual features are the most important in the entire dataset, while LIME can only explain individual prediction decisions. Both techniques are integrated without compromising the classification performance to enhance the transparency, reliability and trustworthiness of the proposed fake news detection framework. Table VIII provides a summary of the explainability techniques used by the proposed framework, their level of analysis, their main findings and their contributions to making the classification of fake news more explainable and interpretable.

TABLE VIII. SUMMARY OF EXPLAINABILITY ANALYSIS

Explainability Technique	**Analysis Level**	**Main Findings**	**Purpose**
SHAP	Global	Identified the most influential textual features contributing to fake and real news classification.	Global model interpretation
LIME	Local	Explained the contribution of individual words for each prediction.	Instance-level interpretation
SHAP + LIME	Combined	Provided complementary global and local explanations that improve transparency and user confidence.	Comprehensive explainability

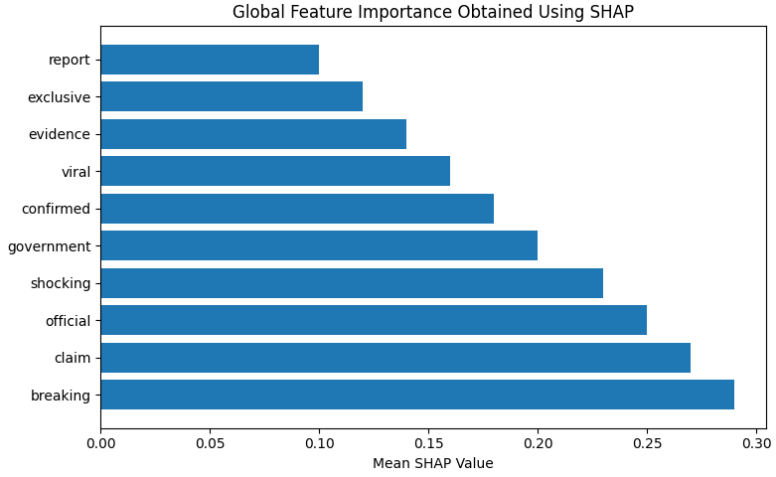

Fig. 9. Global Feature Importance Obtained Using SHAP

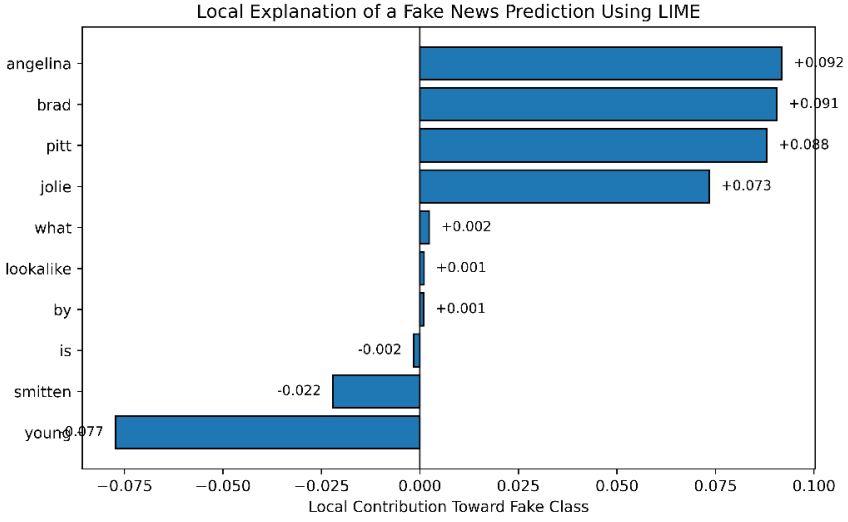

Fig. 10. Local Explanation of a Fake News Prediction Using LIME.

4.5 Comparative Analysis

The performance of the proposed explainable fake news detection framework was assessed and compared with other recent and state-of-the-art fake news detection algorithms reported in the literature to obtain a comprehensive analysis of the effectiveness of the framework. Both quantitative and qualitative classification performance and model interpretability are taken into account. The proposed framework integrates an efficient TF-IDF feature representation, the Linear Support Vector Machine (Linear SVM) classifier and complimentary Explainable Artificial Intelligence (XAI) techniques, supporting reliable classification and transparent decision interpretation, while most of the existing studies focus on the predictive accuracy.

The studies on detecting fake news can be broadly classified into the conventional machine learning methods, transformer-based language models, and explainable artificial intelligence (XAI) methods. Transformer-based models like RoBERTa or XLM-R have achieved state-of-the-art results in contextual language understanding and have performed extremely well on the prediction task but are typically resource-intensive and are black-box models. Alternatively, standard machine learning approaches using TF-IDF and Linear SVM are still appealing due to their computational efficiency, effectiveness on high-dimensional feature spaces and ease of deployment. The use of SHAP and LIME in the proposed framework further strengthens it by adding other explanations that are complementary to the original framework, both at global and local levels. Table IX compares the proposed framework with some representative fake news detection methods that are based on machine learning, transformer, and explainable AI. The comparison not only reveals the predictive performance of each model, but it also shows the explainability features they offer.

TABLE IX. COMPARATIVE ANALYSIS WITH RECENT FAKE NEWS DETECTION STUDIES

Reference	Model	Dataset	Explainability	Accuracy (%)	F1-score (%)
[22]	TF-IDF + Linear SVM	LIAR	No	77.80	84.90
[23]	TF-IDF + Logistic Regression	FakeNewsNet	No	78.60	85.70
[24]	RoBERTa	Fake News Dataset	No	87.60	86.90
[25]	Adaptive Ensemble	Public Fake News Dataset	SHAP + LIME	99.40	99.50
[26]	HEMT-Fake (XLM-R)	Multilingual Fake News Dataset	SHAP + LIME	89.00	89.00
Proposed Framework	**TF-IDF + Linear SVM**	**FakeNewsNet (GossipCop + PolitiFact)**	**SHAP + LIME**	**79.40**	**87.30**

The comparative results show that the proposed framework constitutes a reasonable trade-off between classification performance, computational efficiency and model interpretability. Transformer-based and ensemble learning models generally has a higher classification accuracy but consumes significantly larger computational resources. The proposed TF-IDF and Linear SVM framework, on the other hand, obtained an F1 score of 87.30% and a recall score of 95.34%, which proves the high performance of it in terms of detecting fake news while guaranteeing low computational complexity. Moreover, the combination of SHAP and LIME yields not only global but also local explanations of classification decisions, making the proposed framework relevant and applicable in practice where interpretability and the deployment efficiency of the model are also critical.

5. DISCUSSION

The experimental results illustrate that the proposed framework for detecting fake news based on the concept of explainable machine learning and explainable AI (XAI) can effectively detect fake news and provide transparency of the model while still maintaining a high level of accuracy. The proposed framework demonstrated high accuracy of 79.40%, F1-score of 87.30%, and recall of 95.34%, showing promising performance in detecting fake news articles. The high recall is important in misinformation detection because it decreases the number of fake news articles that get mistakenly marked as real news. The combination of SHAP and LIME greatly improves the interpretability of the proposed framework. While LIME gives explanations of the individual prediction decisions by identifying the salient words that affected the classification outcome, SHAP gives global explanations by identifying the features that are most salient in the classification process for the entire dataset. This explainability feature is complementary, providing the user with an understanding of the overall behavior of the classifier and the "how" component of predictions, giving greater confidence in automated decision making.

The proposed framework provides a desirable balance between interpretability, computational cost, and prediction accuracy, which is advantageous compared with recent transformer-based methods and ensemble learning. Several deep learning models demonstrated high classification accuracy but most of these models require significantly higher computational resources and are considered as black-box systems. Different from this, the proposed Linear SVM-based framework is significantly less costly in terms of computational complexity but achieves competitive performance in fake news detection and also allows to explain the obtained results in terms of SHAP and LIME.

Several limitations of the experimental evaluation are also noted. The framework was assessed by applying it to the GossipCop and PolitiFact sub-sets of the FakeNewsNet that consist mainly of English-language news articles and might not be fully representative of new patterns in misinformation or multilingual news. Furthermore, transformer-based

language models can learn from long textual documents to identify deep contextual semantic relationships, which the use of TF-IDF does not enable the model to learn. However, the achieved results show that the proposed approach allows for an effective and interpretable solution for the practical detection of fake news, especially when computational resources are limited.

In general, the proposed framework shows that endowing the traditional text representation methods with the ability of explainable machine learning provides a feasible balance between classification performance, computational efficiency, and interpretability. TF-IDF, Linear SVM, SHAP, and LIME provide a transparent framework for detecting fake news that can be implemented in digital journalism, social media monitoring, fact-checking systems, and other decision-support applications that demand accurate and interpretable AI.

6. CONCLUSION

This paper proposed an explainable fake news detection framework which combines Textual Feature Extraction based on TF-IDF with the Linear Support Vector Machine (Linear SVM) classifier and the explainability techniques SHAP and LIME. The proposed structure is efficient, reliable and comprehensible, and it allows the accurate detection of fake news along with detailed clarifications of the prediction results. In contrast to other existing fake news detection methods that have mostly concentrated on classification performance, the proposed method not only focuses on prediction performance but also incorporates model interpretability to boost user trust in automated decision making. The experimental study was done on the GossipCop and PolitiFact benchmark subsets of FakeNewsNet repository. The results obtained showed that the proposed framework has obtained an accuracy of 79.40%, precision of 80.51%, recall of 95.34%, F1-score of 87.30% and ROC-AUC of 81.40%. The results suggest that the proposed model is very effective in detecting fake news articles while being able to classify overall articles in a balanced way. Furthermore, the incorporation of SHAP and LIME improved the transparency of the proposed framework by not only offering explanations for the predictions, but also by giving a global feature importance analysis. In this sense, the proposed model not only yields reliable classification results, but it also allows the user to learn the reasons for each classification, which makes it applicable for practical use, including digital journalism, automatic fact-checking or social media content moderation. Moving forward, research will be conducted on the representation of language in context to enhance classification accuracy, on multilingual fake news detection, on the integration of multimodal data (textual and visual data) and on compact explainable deep learning architectures for real-time deployment in large-scale digital media platforms.

Funding:

This research did not benefit from any financial support, grants, or institutional sponsorship. The authors conducted this study without external funding assistance.

Conflicts of Interest:

The authors declare that they have no conflicting interests.

Acknowledgment:

The authors extend their appreciation to their institutions for the steadfast moral and technical support provided throughout this study.

References

[1] S. V. Balshetwar, A. Rs, and D. J. R, "Fake news detection in social media based on sentiment analysis using classifier techniques," Multimedia Tools and Applications, vol. 82, no. 23, pp. 35781–35811, 2023.
[2] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," Social Network Analysis and Mining, vol. 13, no. 1, Art. no. 30, 2023.
[3] S. Rastogi and D. Bansal, "A review on fake news detection 3T's: Typology, time of detection, taxonomies," International Journal of Information Security, vol. 22, no. 1, pp. 177–212, 2023.
[4] H. Zhou, "Research of text classification based on TF-IDF and CNN-LSTM," in Proc. Journal of Physics: Conference Series, vol. 2171, no. 1. Bristol, U.K.: IOP Publishing, Art. no. 012021, 2022.
[5] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development," Clinical and Translational Science, vol. 17, no. 11, Art. no. e70056, 2024.
[6] S. R. A. Parisineni and M. Pal, "Enhancing trust and interpretability of complex machine learning models using local interpretable model-agnostic SHAP explanations," International Journal of Data Science and Analytics, vol. 18, no. 4, pp. 457–466, 2024.
[7] A. Saeed and E. Al Solami, "Fake news detection using machine learning and deep learning methods," Computers, Materials & Continua, vol. 77, no. 2, 2023.
[8] G. Corsi, "Evaluating Twitter's algorithmic amplification of low-credibility content: An observational study," EPJ Data Science, vol. 13, no. 1, pp. 1–15, 2024.
[9] B. Palani, S. Elango, and V. Viswanathan K., "CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT," Multimedia Tools and Applications, vol. 81, no. 4, pp. 5587–5620, 2022.
[10] S. Tufchi, A. Yadav, and T. Ahmed, "A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities," International Journal of Multimedia Information Retrieval, vol. 12, no. 2, Art. no. 28, 2023.

[11] Y. Li, H. He, J. Bai, and D. Wen, "MCFEND: A multi-source benchmark dataset for Chinese fake news detection," in Proc. ACM Web Conference 2024, pp. 4018–4027, 2024.
[12] L. Mishchenko and I. Klymenko, "Method for detecting fake news through writing style," 2023.
[13] E. Kotei and R. Thirunavukarasu, "A systematic review of transformer-based pre-trained language models through self-supervised learning," Information, vol. 14, no. 3, Art. no. 187, 2023.
[14] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, et al., "Atlas: Few-shot learning with retrieval augmented language models," Journal of Machine Learning Research, vol. 24, no. 251, pp. 1–43, 2023.
[15] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "Explainable AI for retinoblastoma diagnosis: Interpreting deep learning models with LIME and SHAP," Diagnostics, vol. 13, no. 11, Art. no. 1932, 2023.
[16] P. Knab, S. Marton, U. Schlegel, and C. Bartelt, "Which LIME should I trust? Concepts, challenges, and solutions," in Proc. World Conference on Explainable Artificial Intelligence. Cham, Switzerland: Springer Nature Switzerland, pp. 28–52, 2025.
[17] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," IEEE Access, vol. 11, pp. 78994–79015, 2023.
[18] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, et al., "Interpreting black-box models: A review on explainable artificial intelligence," Cognitive Computation, vol. 16, no. 1, pp. 45–74, 2024.
[19] P. Saikia, K. Gundale, A. Jain, D. Jadeja, H. Patel, and M. Roy, "Modelling social context for fake news detection: A graph neural network based approach," in Proc. 2022 Int. Joint Conf. Neural Networks (IJCNN), pp. 1–8, 2022.
[20] Y. Shen, Q. Liu, N. Guo, J. Yuan, and Y. Yang, "Fake news detection on social networks: A survey," Applied Sciences, vol. 13, no. 21, Art. no. 11877, 2023.
[21] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.
[22] Q. A. Hameed, "Leveraging traditional machine learning and TF-IDF for robust fake news detection," International Journal of Computational & Electronic Aspects in Engineering (IJCEAE), vol. 6, no. 3, 2025.
[23] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Investigating response behavior through TF-IDF and Word2Vec text analysis: A case study of PISA 2012 problem-solving process data," Heliyon, vol. 10, no. 16, 2024.
[24] Y. Blanco-Fernández, J. Otero-Vizoso, A. Gil-Solla, and J. García-Duque, "Enhancing misinformation detection in Spanish language with deep learning: BERT and RoBERTa transformer models," Applied Sciences, vol. 14, no. 21, Art. no. 9729, 2024.
[25] B. Bhattarai, O. C. Granmo, and L. Jiao, "Explainable Tsetlin machine framework for fake news detection with credibility score assessment," in Proc. 13th Language Resources and Evaluation Conf. (LREC), pp. 4894–4903, 2022.
[26] R. Jadhav, V. Meshram, A. Bhosle, K. Patil, S. Dash, and S. Jadhav, "Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1690616, 2025.