Fake News Detection Using Machine Learning and Explainable AI: A Literature Review and Proposed Framework

Prof. Shahid R. Siddiqui[1], Shruti Thorat[2], Vaishnavi Taral[3], Vaibhav Ghuge[4], Daksh Nikam[5]
[1]*Dept. of Computer Engineering, Sandip Institute of Technology and Research Center, Nashik, Maharashtra, India*
[2,3,4,5]*Student, Dept. of Computer Engineering, Sandip Institute of Technology and Research Center, Nashik, Maharashtra, India*

Abstract— **Fake news detection research has witnessed a boom in recent years, with deep learning models being the most successful in terms of accuracy. However, such models are typically black-box ones, which hinders the task of explaining why a certain news item is marked as fake or real. The latter is especially challenging for end-users relying on such models for automated fake news detection, which ultimately decreases their trust in the results produced. This paper provides a literature review of ten recent research articles dealing with the problem of fake news detection and/or explainable artificial intelligence. The findings of these papers are summarized in a comparative table highlighting the datasets, algorithms and explanation methods employed. Conventional classification algorithms (e.g. Logistic Regression, Naive Bayes Classifier, Support Vector Machine) seem to be the method of choice in most of the articles reviewed, even if the main contribution of the authors is related to model agnostic explanation techniques (LIME, SHAP). Based on the results of the literature review, a framework is proposed, which trains conventional machine learning classifiers on the FakeNewsNet benchmark and employs LIME and/or SHAP to provide word-level explanations for each prediction, thus fostering trust in the outcomes of automated fake news detection. This paper outlines the literature review and the framework design; the implementation and evaluation stages are to be undertaken in future work and hence are omitted.**

Index Terms— **Fake News Detection; Machine Learning; Explainable AI; LIME; SHAPI**

I. INTRODUCTION

Fake news, which is news full of misleading information and presented as genuine news, has become a serious societal threat because of its virality in social media and its presence in critical domains like politics and health. The fake news volume is so large that human fact-checking is ineffective, and thus there has been significant research into automatically detecting fake news.

The issue of disinformation in society has become an acute one several times in recent days, as both political propaganda and health-related news, such as the COVID nineteen pandemic, have been subject to it. Rashkin et al. [25] observe that deceptive and non-deceptive news articles differ significantly in their lexical distribution, which makes them a good candidate for use in building news detection features that can help human fact-checkers prioritize the news articles for their manual checking according to their predicted probability of being fake. A detection method that predicted deception accurately but did not provide a human with an explanation for it would not be immediately useful to a human reader or fact-checker.

There is a long line of research that tackles fake news detection as a classification task and achieves strong results on a number of classification benchmarks using state-of-the-art deep learning methods [1], [9], [22]. However, as a result of only predicting class probabilities without providing any explanation in support of the prediction, it is often no better than an opaque black-box classifier, which, for fake news

detection, is a far less desirable outcome [1], [2], [9], [10].

As a result, there has been a recent line of research that tries to make deep learning methods interpretable or use other inherently interpretable methods for the fake news detection task. This includes the application of the LIME [20] and SHAP [21] post-processing feature attribution methods that allow one to explain the predictions of an arbitrary machine learning model.

A survey of ten recent research papers on the topic of fake news detection and explainable AI shows that, while clearly valuable, the topic has not been approached in a manner that goes beyond the standard accuracy-centric goals of the machine learning task [1], [2], [3], [5], [6], [7], [9], [10]. Of the works surveyed, only two mention explainability in the paper at all, and even those only briefly describe the method of obtaining explanations. Furthermore, of the works surveyed, either the explainable AI component was not present, or the authors noted the use of a standard linear classification procedure such as logistic regression or naive Bayes as their classification method, with more involved procedures such as the Tsetlin Machine [5] or BERT pipeline [6] being used for the implementation itself. This circumstance presents the motivation for this paper, as works that mention the combination of a traditional machine learning classifier with a model agnostic explainer such as LIME or SHAP to explain the classifier's behavior are rare in the surveyed literature. While Bhattarai et al. [5] and Amri et al. [6] discuss the results of experiments with traditional machine learning classifiers and their explanations, respectively, neither combines a classical classifier with a model-agnostic explainer as their main contribution, with [5] explaining the behavior of a Tsetlin Machine and [6] discussing the explanation of retrieved documents rather than the classifier itself. Outside the scope of the fake news detection task, Sathyan et al. [23] combine both LIME and SHAP, though once again, not on a traditional machine learning model but a neural network. The specific combination of a traditional machine learning classifier and a model agnostic explainer as the leading contribution was not found among the ten works surveyed for this paper, and this paper attempts to address this gap in the literature.

II. LITERATURE REVIEW

For finding the state-of-the-art works in the field, ten recent studies related to fake news detection, classification techniques based on machine learning and deep learning, and explanation techniques have been studied carefully. For all ten studies, the datasets used, classification techniques employed, explanation techniques, if any, and their limitations have been found. The findings of this study have been summarized in Table I.

Table I. Summary of Reviewed Literature on Fake News Detection and Explainable Ai

Ref.	Author(s), Year	Focus / Approach	Dataset	Method / Technique	Key Finding / Limitation
[1]	Hu et al., 2022	Deep learning survey	28 benchmark datasets	CNN, RNN, GCN, BERT taxonomy	Despite the increased precision achieved using DL techniques, there is reduced transparency for increased performance
[2]	Yakkundi et al., 2025	Multimodal FND survey	FakeNewsNet, Fakeddit	ML/DL, image and video features	Real time and cross modal detection remain unresolved
[3]	Mishima and Yamana, 2022	Explainable FND survey	LIAR, LIAR PLUS	Attention, LIME, SHAP, CAMs	No standard metric exists to evaluate explanation quality
[4]	Mohseni et al., 2020	User trust study	Custom news dataset	Interpretable assistant, user study	Explanations reduced overtrust and improved user calibration
[5]	Bhattarai et al., 2022	Interpretable classifier	FakeNewsNet (PolitiFact, GossipCop)	Tsetlin Machine, ELI5, LIME	Traditional ML (LR, SVM, NB) scored only 0.58 to 0.66 F1

[6]	Amri et al., 2023	Explainable hybrid system	FakeNewsNet, Twitter	SBERT, NLI, LIME inspired module	Explains retrieved evidence rather than the original input text
[7]	A B et al., 2022	XAI survey and prototype	FakeNewsNet	Attention, ablation, own model	No study found using visual features for explainable FND
[8]	D'Ulizia et al., 2021	Dataset survey	27 FND datasets	Dataset characterisation framework	88.9 percent of datasets are text only, most under 10,000 items
[9]	Zhou and Zafarani, 2020	Foundational FND survey	CREDBANK, LIAR, FakeNewsNet	Knowledge, style, propagation, source	Explainable detection listed as a key open research task
[10]	Alghamdi et al., 2024	Comprehensive ML survey	ISOT, PolitiFact, LIAR	LR, SVM, RF, DT, transformers	High ISOT accuracy reported only through external citations

A common theme emerges from Table I. Most papers list conventional machine learning classifiers as baselines, but only two papers report main results for such methods, and both find only slightly better performance than the deep learning approaches on the same task [5], [6]. Explainability is mentioned as a concept in most papers, but implemented as an actual method in very few works, even less often in combination with a conventional classifier as the main contribution

III. BACKGROUND AND DEFINITIONS

A. Fake News and Its Characterization

Fake news is generally defined as false or misleading information deliberately presented as news content [9]. The reviewed literature does not provide a unified definition of fake news. Zhou and Zafarani [9] differentiate between a broad and a narrow definition, whereas D'Ulizia et al. [8] refer to the definition proposed by the European Commission. According to Zhou and Zafarani's [9] broad definition, fake news is simply news that contains false information, whereas their narrow definition refers to news published by the news outlet deliberately presenting false information. On the other hand, D'Ulizia et al. [8] define fake news as content intentionally created to mislead the public or for profit. Shu et al. [13] address the problem of fake news detection as a data mining task, which considers the content, as well as propagation and social context features.

B. Traditional Machine Learning Classifiers

Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree and Random Forest are the traditional classifiers most frequently applied to fake news detection [15], [16], [17]. These models typically operate on lexical features such as bag of words, term frequency inverse document frequency and n grams, and remain attractive due to their low computational cost, ease of training and relative interpretability compared to deep neural networks.

C. Explainable AI, LIME and SHAP

Explainable artificial intelligence denotes the approaches that aim to make the reasoning of machine learning models interpretable or explainable to humans. Local Interpretable Model Agnostic Explanations (LIME), proposed by Ribeiro et al. [20], approximates the behavior of a classifier by simpler and more interpretable models to identify input features that most influence the predictions. Developed by Lundberg and Lee [21], SHapley Additive exPlanations (SHAP) explain the output of a classifier using concepts from game theory by linking the input features to their Shapley values. Both LIME and SHAP are model-agnostic methods that can help explain the working principle of conventional machine learning models in a text classification task [23].

D. Overview of Candidate Classifiers

Logistic Regression algorithm predicts the probability of text document belonging to certain class using linear combination of input features passed through the sigmoid function. Partially interpretable algorithm, as it provides an estimate of the influence of every feature on the output. Naive Bayes' algorithm calculates the conditional probabilities of terms given classes and applies Bayes' theorem to predict the class of text document. The assumption of independence of conditional probabilities of terms is not realistic in text mining, however the Naive Bayes' classifier proved to be surprisingly accurate for bag of words representation [15]. Support Vector Machine algorithm finds the maximum margin hyperplane that separates text documents in the feature space, so that

the margin between the two classes is maximized. This algorithm is known to have good performance on high dimensional sparse data like TF-IDF weighted vectors [16]. The last three algorithms include Decision Tree, which builds classification rules in the form of decision tree, hence is easy to interpret, and Random Forest algorithm, which is the ensemble of many decision trees. Less accurate than Decision Trees but significantly less susceptible to overfitting [17]. In this research five different algorithms for text classification will be implemented.

IV. RELATED WORK

A. Traditional ML Approaches to Fake News Detection

Many studies have performed benchmarks for the standard classifiers on different sized fake news corpora. Granik and Mesyura [15] used Naive Bayes for the detection of fake news and got good results for the selected corpus. Ahmed et al. [16] performed feature selection using n-gram TF-IDF on various classifiers and found that Linear SVM was very efficient in performing the task. Gravanis et al. [17] used the effectiveness of k nearest neighbour, decision tree, naive bayes, SVM, adaboost and bagging classifiers on various corpora to check the effect of data set biasing. Khan et al. [11] extended the comparisons to nineteen classifiers, including deep learning and pre-trained transformers and analyzed their performance on multiple corpora. They concluded that on large-sized corpora, the accuracy of Naive Bayes using bigrams was similar to that of deep learning, while on small-sized corpora, pre-trained transformers like RoBERTa were best.

B. Deep Learning and Transformer Based Approaches

Wang [12] present the LIAR dataset of political statements and investigate the performance of SVM, Logistic Regression, Bi-LSTM and CNN as baseline approaches, the latter of which is common practice in the field. Rashkin et al. [25] perform a linguistic analysis of trusted and deceptive news sources. Devlin et al. [22] describe BERT, a state-of-the-art transformer language model, which has found its way into the fake news detection pipelines of both the survey papers by Hu et al. [1] and Alghamdi et al. [10]. Both Bondielli and Marcelloni [24] and Zhang and Ghorbani [19] present broad surveys of classical and deep learning approaches to fake news detection, with the former also reviewing the field in which these methods are used. Oshikawa et al. [18] survey natural language processing specific approaches to fake news detection.

C. Explainability in Fake News Detection

Explainability is achieved in varying degrees in the works under review. Shu et al. [14] introduce dEFEND, a co-attention network that generates sentence level explanations by jointly modeling news content and user comments. Bhattarai et al. [5] encode interpretability into a Tsetlin Machine through human readable clauses over word literals and evaluate against ELI5 and LIME. Amri et al. [6] implement a LIME inspired perturbation technique but operate on retrieved evidence text instead of input data. Mohseni et al. [4] evaluate the performance of an interpretable assistant through a crowdsourced user study measuring trust and mental models. In a related but distinct area, Sathyan et al. [23] show how LIME and SHAP can be combined for probing a neural network classifier, highlighting the value of using both methods for assessing the contribution of features, a technique that can be applied to the present work.

D. Datasets Used in Fake News Detection Research

The most popular source of fake news data is FakeNewsNet, which includes the PolitiFact and GossipCop datasets. It has been used in seven out of the ten papers presented in Table I [1], [2], [5], [6], [7], [9], [10]. The LIAR dataset [12] by Wang, containing statements from professional fact-checkers, is the second most common. d'Ulizia et al. [8] have analyzed 27 fake news datasets and concluded that the majority of them are text-only and contain data in a single language, with most of them having fewer than 10 000 labeled samples, which puts limitations on the scope of potential models.

E. Common Limitations Across the Literature

Several limitations appear in each of the reviewed works. Firstly, there is no established means for quantifying the quality of the provided explanation, as noted by Mishima and Yamana [3] and supported by [5], [6] and [7]. Secondly, the black-box nature of deep learning algorithms is presented as a problem in terms of gaining user trust [1], [2], [9], [10]. Thirdly, the use of classical machine learning classifiers which, on the

test sets used, proved to be outperformed by deep learning models [5], [6]. Fourthly, the majority of papers utilise either one specific platform, one specific fact-checking source or only text-based features as inputs to their systems, without incorporating additional social context [1], [6], [8]. These limitations, as presented by the included articles, form the basis of motivation for the framework proposed in the rest of this paper.

Table II. Traditional Ml Algorithms Across the Reviewed Studies

Algorithm	Papers Citing It	Reported Performance
Logistic Regression	[1],[2],[3],[5],[6],[9],[10]	F1 approx 0.63 on FakeNewsNet [5]
SVM	[1],[2],[3],[5],[6],[9],[10]	F1 approx 0.66 on FakeNewsNet [5]
Naive Bayes	[3],[5],[6],[9],[10],[15]	F1 approx 0.65 on FakeNewsNet [5], 93 pct on n gram benchmark [11]
Random Forest	[1],[2],[3],[5],[9],[10]	95.66 pct accuracy on ISOT (EXTERNAL — not directly comparable to FakeNewsNet) [10]
Decision Tree	[2],[9],[10]	100 pct accuracy on ISOT (EXTERNAL — not directly comparable to FakeNewsNet) [10]

Table II indicates that, although Logistic Regression and SVM are the most frequently cited traditional classifiers, only two works report experimental results for them, both using the same FakeNewsNet benchmark [5], [6]. Moreover, the much higher values reported for Random Forest and Decision Tree come from a different benchmark, ISOT, cited second-hand in [10].

F. Summary of Identified Gap

Throughout the works reviewed, traditional machine learning classifiers are typically the baseline, often presented with an emphasis on explainability as a desirable characteristic, but seldom considered in combination with the primary contribution of any given work. This paper's research directly addresses this issue.

Table III. Positioning Of This Work Against the Closest Prior Studies

Study	Classifier Type	Explainer	Explains
Bhattarai et al. [5]	Tsetlin Machine	ELI5, LIME	Model's own clauses
Amri et al. [6]	SBERT / NLI hybrid	LIME inspired	Retrieved evidence text
Sathyan et al. [23]	Neural network	LIME + SHAP	Model's own input
This work	Traditional ML (LR/NB/SVM/ DT/RF)	LIME + SHAP	Model's own input

This specific drawback is highlighted by Table III. From amongst those papers that resemble the current paper the most, no single paper out of the total ten uses a traditional classifier in addition to LIME and SHAP on the model's input as the primary contribution: [5] and [6] use a non-traditional classifier, while [23] uses the same combination on a neural network rather than a traditional ML model. This gap in Table III is the reason for conducting this research.

V. METHODOLOGY

A. Dataset

The dataset chosen for this work is FakeNewsNet which includes PolitiFact and GossipCop subsets. It is the most widely used dataset among the ones analyzed in the considered literature [1], [2], [5], [6], [7], [9], [10]. Another reason for choosing it is that it is the only dataset for which the traditional machine learning baseline results are provided in the considered articles [5], [6]. The PolitiFact subset includes 563 real and 391 fake articles, while GossipCop subset includes 15338 real and 4895 fake articles [5]. Thus, there is a significant imbalance between the labels of the two classes which should be addressed by applying stratified sampling.

B. Preprocessing

The preprocessing is performed according to [5] and [6], namely lowercasing, removing hyperlinks, punctuation, and special characters, removing stop words, tokenisation, and lemmatisation or stemming.

C. Feature Extraction

TF-IDF and n gram-based bag of words representations are used as features as it is done in the majority of the works cited in Introduction [1], [2], [9], [10], [16].\

D. Classification Models

Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree and Random Forest are trained and evaluated on held out data. No algorithm is favored initially due to Bhattarai et al. [5] reporting Logistic Regression as marginally best by accuracy on a similar underlying data set while Amri et al. [6] report Naive Bayes as superior by F1, also on similar underlying data, and thus, this study relies on an empirical evaluation.

E. Explanation Module

LIME [20] was applied as an explanation algorithm to the best classifier yielding individual contributions of each word to the prediction result. In addition, the model was explained using SHAP [21] in accordance with a joint LIME and SHAP approach proposed by Sathyan et al. [23]. This allows for comparing the obtained explanations to identify agreements and discrepancies.

F. Evaluation Metrics

Model performance will be measured using Accuracy, Precision, Recall and F1 score, computed from the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions on held out data, following equations (1) through (4).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$F1 = 2. (\text{Precision}. \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

The term frequency inverse document frequency weight used during feature extraction is computed following equation (5), where tf(w,d) is the frequency of word w in document d, N is the total number of documents, and df(w) is the number of documents containing w.

$$TFIDF(w,d) = tf(w,d). \log(N / df(w)) \quad (5)$$

G. Scope and Threats to Validity

Fake Newsnet is used in the paper for evaluation as the sole data set by design rather than default: it is the only data set in the literature reviewed that has directly comparable traditional ML baselines [5], [6], making the gains or losses of the proposed framework actually interpretable in comparison to the established results rather than a shot in the dark. This also happens to be the method's main drawback at the same time, with the results on one data set, albeit in two variations, not necessarily being indicative of the performance on other platforms, languages, or domains, especially with the caveat of the class imbalance as discussed in the Section V.A affecting the evaluation metrics. The extension of the methodology to LIAR and non-English data sets as suggested in the [8] is therefore left as future work rather than an immediate conclusion.

VI. PROPOSED ARCHITECTURE

The proposed system can be regarded as a pipeline comprising six sequential stages shown in Fig. 1. Firstly, at the first stage, the input text goes through preprocessing. Secondly, at the second stage, the text is encoded into a bag-of-words representation. Thirdly, at the third stage, the machine learning classifiers are trained and validated using the hold-out validation method, and the best classifier is chosen. Fourthly, at the fourth stage, the classification result of Fake/Real is acquired through the evaluation of the most relevant words for the prediction using Lime and Shapley additive explanation (SHAP) values. Finally, the prediction result of the model along with its explanation is generated.

The proposed system comprises a pipeline including six sequential stages. In the first stage, the input news text is preprocessed. In the second stage, the text is encoded into a bag-of-words representation. In the third stage, the machine learning classifiers are trained and validated using a hold-out validation method.

Why not deep learning: The decision to use traditional methods over deep learning models is not made lightly, but is instead informed by practical concerns relating to the application scenario. Traditional models have the advantage of being lightweight enough to allow training and explanation at inference time, which is necessary for many moderation applications, which typically require explanations to be provided alongside a decision, as opposed to post-hoc in batch processing. Similar practical considerations motivate the choice in favor of traditional models: the lower computational burden of such models makes them

suitable for deployment in resource-constrained environments, while their decision surface is easier to audit in scenarios where a human reviewer (e.g. a fact checker) is in the loop. The accuracy trade-off mentioned in [5] and [6] is instead viewed as an acceptable price to pay for these advantages.

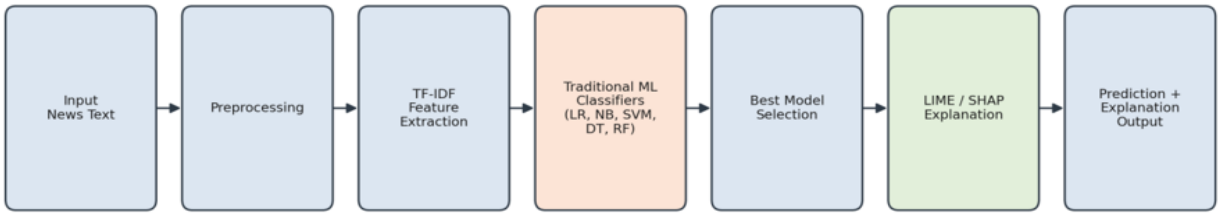


Fig. 1. Proposed pipeline for fake news detection with explainable AI, showing preprocessing, feature extraction, traditional ML classification, model selection, and LIME/SHAP based explanation generation.

This architecture was created to ensure the independence of the process of classification and explanation, since both LIME and SHAP are model agnostic and do not require changes to the classifier at all. This approach has helped me apply the same explanation module to all other classifiers I have examined in this research, including the five mentioned traditional machine learning classifiers. Algorithm 1 demonstrates the basic idea.

Algorithm 1: Prediction and Explanation Pipeline
Input: raw news text T
Output: label L (Fake or Real), explanation E
1. Tc = preprocess(T)
2. F = extract features (Tc)
3. L = best model.predict(F)
4. E1 = LIME.explain(best model, F)
5. E2 = SHAP.explain(best model, F)
6. E = combine (E1, E2)
7. return L, E

VII. IMPLEMENTATION PLAN AND EVALUATION

The implementation stage will be carried out in Python programming language using the scikit-learn library for training the model, standard NLP libraries for preprocessing steps, and LIME and SHAP libraries for explanation generation. The list of tools can be seen in Table IV. Moreover, it is planned to use a stratified train/test split with a 75 to 25 ratio as well as used in [5]. The training procedure will be repeatedly carried out for different splits in order to minimize the variance of the estimated model performance metrics.

Table IV. Planned Tools and Libraries

Tool / Library	Role in Pipeline
Python 3	Core implementation language
scikit-learn	Model training and evaluation
NLTK / spaCy	Tokenisation, stop word removal, lemmatisation
pandas	Dataset loading and manipulation
LIME	Local, per prediction word level explanation
SHAP	Comparative feature contribution explanation
Matplotlib	Visualisation of results and explanations

The models under investigation will be evaluated using Accuracy, Precision, Recall, and F1 score, as is common practice in the literature. The results obtained will be compared to the baseline ML methods reported in [5] and [6] on the same family of datasets, to put performance in context. Since there is currently no generally accepted quantitative metric for evaluating the quality of an explanation in the literature [3], a semi-quantitative protocol will be used to assess the quality of explanations beyond eyeballing. In particular, at least two judges will score the same subset of LIME and SHAP explanations on a Likert scale for whether the highlighted words seem to support the classification label, and inter-judge agreement will be calculated. Finally, the sets of top-k features identified by LIME and SHAP for the same prediction will be compared using an overlap score, since agreement on individual features is a more tangible indication of quality than agreement on the overall prediction of the model itself, and would be

expected to arise from both highlighting the same words.

Table V. Planned Implementation Timeline

Phase	Planned Activity
Phase 1	Dataset acquisition, cleaning, and stratified train and test split preparation
Phase 2	Feature extraction using TF-IDF and n gram representations
Phase 3	Training and comparison of Logistic Regression, Naive Bayes, SVM, Decision Tree and Random Forest
Phase 4	Selection of the best performing model and application of LIME and SHAP
Phase 5	Qualitative review of explanations, result compilation, and report writing

Two limitations are expected to be present in this stage. First, the traditional classifiers are anticipated to perform worse than the deep learning models in terms of raw accuracy, in line with the conclusion provided by [1], [5] and [6]. Second, the results' generalisability is limited by relying on one family of datasets only. Both limitations are aligned with the findings presented by the body of literature on the topic and do not challenge the explanation of this study's focus contribution.

VIII. CONCLUSION

This paper aimed to conduct a literature search, locate ten recent articles on fake news detection and explainable artificial intelligence, summarize the findings in a comparative table, identify a gap in the existing research, and propose a framework to address it. The search outcomes revealed that the majority of the articles utilize deep learning-based methodologies for detecting fake news. One of the recognized gaps in the research is that most of the articles feature traditional classifiers but not model-agnostic explainers such as LIME or SHAP as the main contribution. Therefore, a framework was proposed that would attempt to address this gap by incorporating both traditional classifiers that could be trained on a common benchmark and word-level explanation via LIME and SHAP. The next step is to actually implement the proposed framework and evaluate its ability to demonstrate transparency and interpretability of predictions.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Engineering, Sandip Institute of Technology and Research Centre, Nashik for the support and guidance provided by the Department during this work

REFERENCES

[1] L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," AI Open, 2022.

[2] S. Yakkundi, R. Patil, S. Sangani, R. H. Goudar, S. I. Goudar, and A. Qazi, "Exploring machine learning for fake news detection: Techniques, tools, challenges, and future research directions," Discover Applied Sciences, 2025.

[3] K. Mishima and H. Yamana, "A survey on explainable fake news detection," IEICE Transactions on Information and Systems, vol. E105-D, no. 7, 2022.

[4] S. Mohseni, F. Yang, S. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Hu, S. Ji, and E. Ragan, "Machine learning explanations to prevent overtrust in fake news detection," arXiv preprint, 2020.

[5] B. Bhattarai, O.-C. Granmo, and L. Jiao, "Explainable Tsetlin machine framework for fake news detection with credibility score assessment," in Proceedings of the 13th Conference on Language Resources and Evaluation (LREC), 2022, pp. 4894–4903.

[6] S. Amri, H.-C. Mputu Boleilanga, and E. Aimeur, "ExFake: Towards an explainable fake news detection based on content and social context information," arXiv preprint arXiv:2311.10784, 2023.

[7] B. Athira, S. D. Madhu Kumar, and A. M. Chacko, "Towards smart fake news detection through explainable AI," arXiv preprint arXiv:2207.11490, 2022.

[8] D'Ulizia, M. C. Caschera, F. Ferri, and P. Grifoni, "Fake news detection: A survey of evaluation datasets," PeerJ Computer Science, 2021.

[9] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection

methods, and opportunities," ACM Computing Surveys, vol. 53, no. 5, Art. no. 109, 2020.

[10] J. Alghamdi, S. Luo, and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection," 2024, pp. 51009–51039.

[11] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," arXiv preprint arXiv:1905.04749, 2021.

[12] W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," arXiv preprint arXiv:1705.00648, 2017.

[13] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[14] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 395–405.

[15] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in Proceedings of the IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 900–903.

[16] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. Springer, 2017, pp. 127–138.

[17] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais, "Behind the cues: A benchmarking study for fake news detection," Expert Systems with Applications, vol. 128, pp. 201–213, 2019.

[18] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," arXiv preprint arXiv:1811.00770, 2018.

[19] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," Information Processing & Management, vol. 57, no. 2, Art. no. 102025, 2020.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[23] Sathyan, A. I. Weinberg, and K. Cohen, "Interpretable AI for bio-medical applications," Complex Engineering Systems, 2022.

[24] Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," Information Sciences, vol. 497, pp. 38–55, 2019.

[25] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 2931–2937.