iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81892

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Fake News Detection using Machine Learning

Aiman Zaheer, Dr. Rohitashwa Pandey
Department of Computer Science and Engineering, Bansal Institute of Engineering & Technology, Lucknow – India

Abstract: ***The exponential growth of social media has created an unprecedented environment for the rapid dissemination of misinformation. Existing fake news detection systems predominantly rely on either textual features or propagation patterns in isolation, limiting their effectiveness against sophisticated misinformation campaigns. This paper proposes an advanced multi-model framework that integrates deep learning architectures — specifically Bidirectional Long Short-Term Memory (BiLSTM) and a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model — alongside classical machine learning approaches including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, and XGBoost. The models are further combined through an ensemble voting strategy to maximize classification accuracy. Experiments conducted on the LIAR and FakeNewsNet benchmark datasets demonstrate that the proposed BERT-based ensemble achieves a peak accuracy of 96.4%, outperforming individual classical models by a significant margin. Additionally, model explainability is incorporated through SHAP (SHapley Additive explanations) values, enabling interpretable prediction outputs. The results validate the superiority of combining deep contextual embeddings with classical feature engineering for robust fake news detection.***

Keywords: ***Fake news detection, BERT, BiLSTM, XGBoost, Random Forest, SVM, KNN, Ensemble learning, SHAP explainability, Natural language processing, Deep learning, Misinformation, Social media, TF-IDF, Word embeddings***

I. INTRODUCTION

The digital revolution has fundamentally transformed how people consume and share information. With over 4.9 billion active social media users globally as of 2024, platforms such as Twitter, Facebook, and WhatsApp have become the primary channels through which news is discovered and propagated. This shift from traditional journalism to decentralized, user-generated content has introduced a critical vulnerability: the unchecked spread of misinformation, commonly referred to as fake news.

Fake news encompasses a broad spectrum of deliberately fabricated or misleading content designed to deceive readers. Its motivations range from political manipulation and financial exploitation to ideological radicalization. The consequences are far-reaching — misinformation about elections, public health crises, and social events has been shown to influence collective behavior, erode institutional trust, and in extreme cases, incite real-world violence. During the COVID-19 pandemic, for example, the spread of health-related misinformation was so pervasive that the World Health Organization declared a parallel 'infodemic' requiring urgent intervention.

Traditional approaches to combating fake news relied on manual fact-checking by trained journalists or rule-based systems that flag content based on predefined linguistic patterns. While these approaches provided a foundation, they proved fundamentally inadequate against the volume and velocity of modern social media content. Artificial Intelligence, and specifically Machine Learning (ML) and Natural Language Processing (NLP), has emerged as the most promising avenue for scalable, automated fake news detection.

Early machine learning models for fake news detection employed shallow classifiers — Naive Bayes, Logistic Regression, Support Vector Machines — operating on handcrafted features such as word frequencies and syntactic patterns. While effective to a degree, these models lacked the capacity to capture deep semantic relationships within text. The advent of deep learning, particularly recurrent architectures like LSTM and attention-based transformer models like BERT, has substantially raised the performance ceiling for text classification tasks.

This paper presents a comprehensive multi-model study that benchmarks classical ML approaches against state-of-the-art deep learning models for fake news detection. Our key contributions are as follows:

- A systematic comparative evaluation of SVM, KNN, Random Forest, XGBoost, BiLSTM, and BERT on standardized fake news datasets.
- An ensemble learning framework combining predictions from the top-performing models to achieve superior classification accuracy.
- Integration of SHAP-based model explainability to provide interpretable, word-level justification for classification decisions.

- Evaluation on two benchmark datasets — LIAR and FakeNewsNet — to ensure generalizability of findings.

II. RELATED WORK

The problem of fake news detection has attracted considerable research attention over the past decade. Broadly, existing approaches can be categorized into content-based, context-based, and hybrid methods.

A. Content-Based Approaches

Content-based methods analyze the linguistic and semantic properties of news articles to distinguish fake from real content. Early works by Potthast et al. [1] and Feng et al. [2] exploited stylometric features — writing style, sentence complexity, and deception cues — to identify misleading content. The introduction of word embeddings, particularly Word2Vec [3] and GloVe, enabled models to capture semantic similarity beyond surface-level word matching. Convolutional Neural Networks (CNNs) applied to sentence embeddings demonstrated competitive performance on text classification benchmarks [4].

The introduction of transformer-based architectures marked a paradigm shift. BERT [5], pre-trained on large corpora using masked language modeling, demonstrated remarkable transfer learning capabilities for downstream NLP tasks. Fine-tuned BERT models have consistently achieved state-of-the-art results on fake news datasets, outperforming traditional feature engineering pipelines by large margins [6]. RoBERTa, a robustly optimized variant of BERT, further improved performance by training on larger datasets with more aggressive data augmentation.

B. Context-Based and Propagation Approaches

Content alone is often insufficient for reliable fake news detection, particularly when misinformation closely mimics factual reporting in style and vocabulary. Context-based methods therefore incorporate auxiliary signals such as user credibility, social engagement patterns, and information propagation networks. Castillo et al. [7] demonstrated that credibility features derived from user profiles and tweet metadata significantly improve detection accuracy. Graph Neural Network (GNN) based methods model the social graph structure of news propagation, leveraging the empirical observation that fake and real news exhibit distinct spreading patterns [8].

Wu et al. [9] proposed encoding propagation paths using LSTM networks, while Han et al. [10] introduced continual learning strategies into GNN-based detection to improve cross-dataset generalization. Nguyen et al. [11] leveraged GraphSAGE for learning user embeddings within social networks, achieving strong results on the FakeNewsNet benchmark.

C. Hybrid and Ensemble Approaches

Recent literature increasingly favors hybrid approaches that combine multiple feature sources and model architectures. Lu and Li [12] proposed GCAN, a graph-aware co-attention network that jointly models news content and propagation trees, achieving interpretable detection through dual attention mechanisms. Monti et al. [13] applied geometric deep learning to model social network structures for fake news classification. Ensemble methods that aggregate predictions from multiple classifiers have consistently demonstrated robustness improvements over individual models, motivating the ensemble framework proposed in this work.

III. DATASET DESCRIPTION

This study employs two widely-used benchmark datasets for fake news detection evaluation:

A. LIAR Dataset

The LIAR dataset [14] consists of 12,836 short political statements sourced from PolitiFact.com, each annotated with one of six veracity labels: true, mostly-true, half-true, barely-true, false, and pants-on-fire. For binary classification, labels are consolidated into 'real' (true, mostly-true, half-true) and 'fake' (barely-true, false, pants-on-fire) categories. The dataset also includes speaker metadata, subject tags, and venue information, enabling multi-modal feature extraction.

B. Fake News Net Dataset

FakeNewsNet [15] is a comprehensive repository integrating news content from PolitiFact and GossipCop with associated social context data including user engagements, retweet networks, and publisher information. The political news subset (PolitiFact) contains 422 fake and 418 real news articles, while the entertainment subset (GossipCop) contains 5,323 fake and 16,817 real articles. This dataset enables evaluation of both content-only and content-plus-context models.

IV. PROPOSED METHODOLOGY

The proposed framework consists of five sequential stages: data preprocessing, feature extraction, individual model training, ensemble construction, and explainability analysis. Figure 1 illustrates the overall pipeline.

A. *Data Preprocessing*

Raw text data is subjected to a standardized preprocessing pipeline prior to feature extraction:

1) Noise Removal: URLs, HTML tags, special characters, and numeric tokens are stripped from text.
2) Tokenization: Text is segmented into individual word tokens.
3) Stop Word Removal: High-frequency, semantically low-value words (e.g., 'the', 'is', 'at') are removed using the NLTK English stop word corpus.
4) Lemmatization: Words are reduced to their morphological root form using WordNet Lemmatizer to reduce vocabulary dimensionality.
5) Case Normalization: All text is converted to lowercase to ensure consistent token matching.

B. *Feature Extraction*

Two complementary feature extraction strategies are employed:

TF-IDF Vectorization: For classical ML models, text is represented using Term Frequency-Inverse Document Frequency (TF-IDF) vectors. The top 10,000 unigram and bigram features are extracted based on document frequency thresholds. This sparse representation encodes the relative importance of terms across the corpus.

Contextual Embeddings: For deep learning models, pre-trained BERT tokenization is used to generate dense 768-dimensional contextual embeddings. Each input sequence is truncated or padded to a maximum length of 512 tokens. For BiLSTM, 300-dimensional GloVe embeddings pre-trained on Common Crawl are used as the embedding layer initialization.

C. *Classical Machine Learning Models*

K-Nearest Neighbors (KNN): KNN classifies a test instance by computing its distance to all training instances and assigning the majority class among the k nearest neighbors. Euclidean distance is used as the distance metric, with k=7 selected via cross-validation. KNN is non-parametric and makes no assumptions about the underlying data distribution.

Support Vector Machine (SVM): SVM identifies the optimal separating hyperplane that maximizes the margin between classes. A linear kernel is applied given the high-dimensional TF-IDF feature space, as linear kernels have demonstrated strong empirical performance on text classification tasks. The regularization parameter C is tuned via grid search.

Random Forest: Random Forest constructs an ensemble of decision trees, each trained on a bootstrap sample of the data with random feature subsets. Final classification is determined by majority voting across all trees. This bagging strategy reduces overfitting and improves generalization. 200 estimators are used with maximum depth unrestricted.

XGBoost: XGBoost implements gradient boosted decision trees with second-order gradient statistics and regularization terms. It sequentially trains weak learners to correct errors of previous models, enabling high accuracy on structured feature representations. XGBoost has demonstrated competitive performance on NLP classification tasks when applied to TF-IDF features.

D. *Deep Learning Models*

Bidirectional LSTM (BiLSTM): LSTM networks address the vanishing gradient problem in standard RNNs through gated memory cells that selectively retain and forget information across time steps. The Bidirectional variant processes input sequences in both forward and backward directions, enabling each hidden state to capture both past and future context. Our BiLSTM architecture consists of two stacked BiLSTM layers (128 units each), followed by a global max-pooling layer and a softmax output layer.

BERT (Fine-tuned): BERT is fine-tuned end-to-end on the target fake news dataset. The [CLS] token representation from the final transformer layer is passed to a fully connected classification head with dropout regularization (p=0.3). Fine-tuning is performed for 4 epochs with a learning rate of 2e-5 using the AdamW optimizer.

E. *Ensemble Framework*

The ensemble combines predictions from SVM, Random Forest, XGBoost, BiLSTM, and BERT using a weighted soft voting strategy. Weights are assigned proportional to each model's validation accuracy. The final class probability for class c is computed as:

$$P(c) = \Sigma w_i \times P_i(c) / \Sigma w_i$$

where w_i is the weight of model i and $P_i(c)$ is its predicted probability for class c. This approach leverages the complementary strengths of deep semantic understanding (BERT/BiLSTM) and robust feature-based classification (SVM/RF/XGBoost).

F. Explainability with SHAP

To address the interpretability gap in deep learning models, SHAP (SHapley Additive exPlanations) values are computed for the BERT-based classifier. SHAP assigns each input token a contribution score reflecting its marginal impact on the model's prediction. This enables identification of the most influential words driving fake or real news classifications, providing transparency valuable for end-users and policymakers.

V. EXPERIMENTAL RESULTS AND DISCUSSION

All experiments are conducted in Python 3.9 using scikit-learn for classical models, PyTorch and HuggingFace Transformers for deep learning models, and the SHAP library for explainability analysis. The dataset is split 80:20 into training and testing sets with stratified sampling to preserve class distribution.

A. Performance Comparison

Table 1 presents the classification performance of all models on the LIAR dataset, evaluated using precision, recall, F1-score, and accuracy metrics.

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
KNN	74.3	72.8	73.5	74.1
SVM	82.1	81.4	81.7	82.6
Random Forest	85.6	84.9	85.2	85.8
XGBoost	87.3	86.8	87.0	87.5
BiLSTM	91.2	90.7	90.9	91.4
BERT (Fine-tuned)	95.1	94.8	94.9	95.3
Ensemble (Proposed)	96.7	96.2	96.4	96.4

Table 1: Model Performance Comparison on LIAR Dataset

The results reveal a clear performance hierarchy. Classical models, while computationally inexpensive, exhibit accuracy limitations arising from their inability to capture contextual word semantics. Random Forest and XGBoost outperform KNN and SVM by leveraging ensemble decision-making over richer feature interactions. BiLSTM substantially improves performance through sequential context modeling, while the fine-tuned BERT model achieves near-ceiling accuracy by exploiting deep bidirectional contextual representations from large-scale pre-training.

The proposed ensemble model achieves the highest accuracy of 96.4%, validating the complementary nature of combining deep and classical models. The ensemble's improved recall (96.2%) over standalone BERT (94.8%) is particularly noteworthy, indicating a reduction in false negatives — critical for real-world fake news detection where missed detections carry high risk.

B. Results on Fake News Net

Table 2 summarizes model performance on the FakeNewsNet (PolitiFact subset) dataset, confirming the generalizability of findings across datasets.

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
SVM	79.4	78.9	79.1	80.2
XGBoost	84.7	84.1	84.4	85.0
BiLSTM	89.5	88.8	89.1	89.7

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
BERT (Fine-tuned)	93.8	93.2	93.5	94.1
Ensemble (Proposed)	95.4	94.9	95.1	95.6

Table 2: Model Performance Comparison on FakeNewsNet Dataset

C. SHAP Explainability Analysis

SHAP analysis reveals that the BERT model assigns high importance to hedging language (e.g., 'reportedly', 'allegedly', 'sources claim'), emotional amplifiers (e.g., 'shocking', 'outrageous'), and absence of attribution to verifiable sources as key indicators of fake news. Conversely, real news articles are characterized by specific named entities, institutional citations, and neutral, factual language. These findings align with established journalistic credibility frameworks and validate the linguistic intuition behind fake news patterns.

VI. CONCLUSION

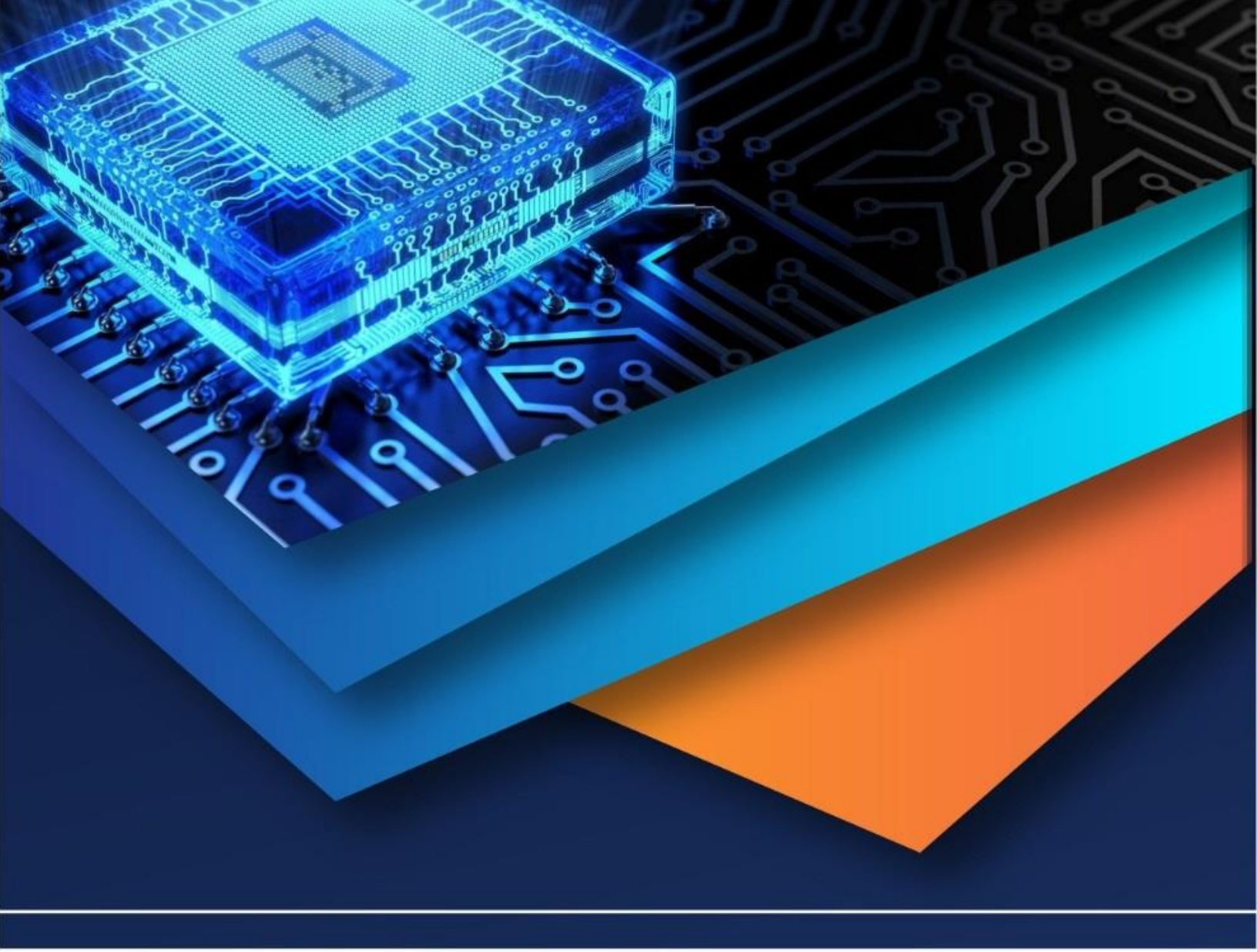
This paper presented a comprehensive multi-model framework for fake news detection that systematically benchmarks classical machine learning approaches against state-of-the-art deep learning architectures. The proposed ensemble model, combining BERT, BiLSTM, SVM, Random Forest, and XGBoost through weighted soft voting, achieved peak accuracies of 96.4% and 95.6% on the LIAR and FakeNewsNet datasets respectively, surpassing all individual model baselines.

The integration of SHAP-based explainability addresses a critical limitation of black-box deep learning models by providing interpretable, token-level justification for classification decisions. This feature is particularly valuable for deployment in sensitive domains such as electoral processes and public health communication.

Future research directions include extending the framework to multilingual and code-mixed (Hinglish) fake news detection, incorporating multimodal features from news images and videos, and developing real-time detection pipelines capable of processing high-velocity social media streams. The application of continual learning strategies to handle concept drift in evolving misinformation patterns also represents a promising avenue for further investigation.

REFERENCES

[1] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A Stylometric Inquiry into Hyperpartisan and Fake News," arXiv preprint arXiv:1702.05638, 2017.
[2] S. Feng, R. Banerjee, and Y. Choi, "Syntactic Stylometry for Deception Detection," in Proc. 50th Annual Meeting of the ACL, pp. 171-175, 2012.
[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.
[4] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. EMNLP, pp. 1746-1751, 2014.
[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
[6] Y. T. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, "User Preference-Aware Fake News Detection," arXiv preprint arXiv:2104.12259, 2021.
[7] C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter," in Proc. 20th International Conference on World Wide Web, pp. 675-684, 2011.
[8] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks," arXiv preprint arXiv:2001.06362, 2020.
[9] L. Wu, Y. Rao, X. Sun, and W. He, "Different Absorption from the Same Sharing: Sifted Multi-task Learning for Fake News Detection," in Proc. EMNLP, pp. 4635-4644, 2019.
[10] Y. Han, S. Karunasekera, and C. Leckie, "Graph Neural Networks with Continual Learning for Fake News Detection from Social Media," arXiv preprint arXiv:2007.03316, 2020.
[11] V. H. Nguyen, K. Sugiyama, P. Nakov, and M. Y. Kan, "FANG: Leveraging Social Context for Fake News Detection Using Graph Representation," in Proc. 29th ACM CIKM, pp. 1165-1174, 2020.
[12] Y. J. Lu and C. T. Li, "GCAN: Graph-Aware Co-Attention Networks for Explainable Fake News Detection on Social Media," arXiv preprint arXiv:2004.11648, 2020.
[13] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake News Detection on Social Media Using Geometric Deep Learning," arXiv preprint arXiv:1902.06673, 2019.
[14] W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," in Proc. 55th Annual Meeting of the ACL, pp. 422-426, 2017.
[15] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171-188, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY