

LIGHTX-FND: A LIGHTWEIGHT EXPLAINABLE TRANSFORMER-DISTILLATION FRAMEWORK FOR ROBUST FAKE NEWS DETECTION ACROSS MULTIPLE DATASETS

Farhan Shah¹, Muhammad Rashid Majeed² and Ayan Ali³

^{1,3}School of Artificial Intelligence, Nanjing University of Information Science and Technology, China

²School of Computer Science, Nanjing University of Information Science and Technology, China

Abstract

The existing fake news detection methods mostly rely on traditional machine learning approaches and text-based classification techniques, but many of these methods face limitations in understanding complex linguistic patterns, achieving high detection accuracy, and providing real-time misinformation verification. This paper proposes an AI-powered Neural Network Fake News Detection Framework (NN-FNDF), a deep learning-based system that leverages natural language processing (NLP) and neural network architectures to automatically identify and classify misleading information. The proposed framework performs comprehensive text preprocessing, including normalization, noise removal, stop-word elimination, and stemming, followed by feature extraction using TF-IDF with unigram and bigram representations to generate an 8,000-dimensional feature space. For classification, the framework introduces a progressive feed-forward neural network architecture with multiple hidden layers (1024, 512, 256, and 128 neurons), batch normalization, staged dropout regularization, and sigmoid-based binary classification. The model is optimized using the AdamW optimizer with adaptive learning rate scheduling and gradient control strategies to improve stability and prevent overfitting. Furthermore, the system incorporates a user-friendly real-time detection interface that provides prediction results, confidence scores, and text analysis information for end users. Experimental evaluation on a dataset containing 6,335 real and fake news articles demonstrates that the proposed framework achieves 92.66% accuracy and a 0.9816 ROC-AUC score, outperforming conventional approaches such as Naive Bayes and K-Nearest Neighbors. The proposed framework provides a scalable and efficient solution for journalists, educators, social media platforms, and fact-checking organizations to combat misinformation. Future work will focus on integrating transformer-based models, explainable artificial intelligence (XAI), multimodal information analysis, and adaptive learning mechanisms to improve robustness against evolving misinformation strategies.

Keywords:

Fake News Detection, Deep Learning, TF-IDF, Neural Networks, Misinformation

1. INTRODUCTION

Nowadays, misinformation and fake news are spreading rapidly in the digital age due to the massive growth of social media platforms and online news networks. False information, misleading narratives, and fabricated content can be easily distributed within seconds, creating serious impacts on public opinion, political decisions, healthcare awareness, financial markets, and social trust [1]. Traditional misinformation detection approaches, including manual fact-checking and rule-based classification methods, are no longer sufficient to handle the increasing volume, complexity, and dynamic nature of online information.

Existing fake news detection techniques mainly depend on traditional machine learning algorithms and manually extracted features, which have limited capability in understanding complex semantic relationships and contextual information within news articles [2]. Although conventional classifiers such as Naive Bayes, Support Vector Machines, and other statistical approaches provide acceptable performance, their effectiveness decreases when dealing with large-scale datasets and evolving misinformation patterns [3]. Recent developments in deep learning have demonstrated significant advantages in automatically learning complex feature representations from textual data, reducing the dependency on manual feature engineering [4]. Deep neural networks have achieved remarkable performance in natural language processing tasks by extracting meaningful patterns from high-dimensional information [5].

For text-based misinformation detection, convolutional neural networks (CNNs) have been successfully applied to capture local semantic features from sentences and documents [6]. Furthermore, recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures have improved sequential text understanding by learning word dependencies and contextual relationships [7], [8]. Word embedding techniques such as Word2Vec have further enhanced language representation by converting textual information into meaningful vector spaces [9]. However, existing approaches still suffer from limitations related to generalization capability, computational efficiency, and limited interpretability of prediction decisions [10].

To overcome these limitations, this paper proposes an AI-powered Neural Network Fake News Detection Framework (NN-FNDF), an intelligent deep learning-based system that integrates natural language processing (NLP), TF-IDF-based feature extraction, and a multilayer neural network architecture for automatic misinformation detection. The proposed framework performs comprehensive text preprocessing, including normalization, noise removal, stop-word elimination, and stemming, followed by TF-IDF feature representation using unigram and bigram information. The extracted features are processed through a deep neural network consisting of multiple hidden layers with batch normalization and dropout regularization to improve learning stability and reduce overfitting.

Furthermore, transformer-based language models such as BERT have recently demonstrated strong contextual understanding capabilities in NLP applications [11]. However, these models often require significant computational resources, motivating the development of lightweight and efficient deep learning solutions for real-time applications. The proposed framework focuses on achieving a balance between detection accuracy, computational efficiency, and practical usability.

The proposed system is experimentally evaluated on a dataset containing 6,335 real and fake news articles. The developed neural network model achieves high classification performance with an accuracy above 92% and a ROC-AUC score of approximately 0.98, outperforming traditional machine learning approaches. The framework also incorporates a user-friendly graphical interface that provides real-time prediction, confidence estimation, and text analysis for practical applications in journalism, education, social media monitoring, and fact-checking systems.

In this paper, NN-FNDF is presented as an effective and scalable solution for automated misinformation detection. Future research will focus on improving model explainability, integrating transformer-based architectures, incorporating multimodal information sources, and developing adaptive learning mechanisms to enhance robustness against continuously evolving misinformation strategies [12]-[15].

2. LITERATURE REVIEW

2.1 OVERVIEW OF FAKE NEWS AND MISINFORMATION TRENDS

The rapid expansion of digital platforms and social media networks has significantly increased the availability and distribution speed of online information. However, this advancement has also accelerated the spread of fake news, rumors, and misleading content, creating serious challenges for information reliability and digital trust. Misinformation has become a global concern due to its influence on political discussions, public perception, financial decisions, and social behavior. Studies have shown that false information can spread faster and reach wider audiences compared with factual information because of user engagement and algorithm-driven content amplification [16].

The increasing complexity of misinformation requires advanced computational approaches capable of understanding linguistic patterns, identifying misleading narratives, and verifying information credibility. Early research focused on analyzing the language characteristics of fake news by examining writing styles, emotional expressions, and semantic patterns. Automatic fake news detection approaches demonstrated that linguistic features and machine learning techniques can effectively distinguish between reliable and unreliable information sources [17]. Furthermore, investigations into political fact-checking revealed that fake news often contains unique language patterns, persuasive structures, and varying levels of truthfulness, making automated verification a challenging research problem [18].

2.2 EXISTING FAKE NEWS DETECTION TECHNIQUES

Traditional fake news detection systems mainly relied on natural language processing (NLP), statistical analysis, and machine learning-based classification approaches. Feature extraction methods such as term frequency analysis, linguistic characteristics, and semantic representations were commonly used to classify news content. Several benchmark studies demonstrated the effectiveness of machine learning models for

fake news identification; however, their performance was highly dependent on manually engineered features and dataset characteristics [19].

With the advancement of deep learning, neural network-based approaches have significantly improved misinformation detection performance. Deep neural networks learn hierarchical feature representations automatically, allowing models to capture complex relationships within textual information. Backpropagation-based neural networks established the foundation for modern deep learning approaches by enabling models to learn effective representations from large-scale data [20]. Furthermore, pretrained language models and generative learning approaches have improved natural language understanding by capturing contextual relationships between words and sentences [21].

Recent research has explored social media-based fake news detection by incorporating user behavior, propagation patterns, and network information. Shu et al. introduced a comprehensive perspective on fake news detection by considering both content-based and social context-based features, highlighting the importance of combining multiple information sources for improved detection accuracy [22]. Similarly, FakeNewsNet provided a large-scale repository containing news content, social context, and temporal information to support advanced misinformation research and evaluation [23].

Deep learning models have also been extended toward multimodal misinformation analysis by combining textual and visual information. The availability of large-scale datasets and improved computer vision techniques has enabled researchers to analyze manipulated images and visual misinformation along with textual content [24]. Convolutional neural network architectures have demonstrated strong capabilities in extracting complex features from large-scale data, making them suitable for automated content analysis applications [25].

2.3 LIMITATIONS OF CURRENT APPROACHES

Although significant progress has been achieved in fake news detection, existing approaches still face several limitations. Traditional machine learning methods often depend on manually extracted features and have limited capability in understanding complex semantic relationships, contextual meaning, and evolving misinformation patterns. Deep learning approaches provide improved feature learning capabilities, but they may require large amounts of labeled data and significant computational resources.

Another major limitation is the lack of model interpretability. Many advanced deep learning models provide classification results without explaining the reasons behind their decisions, which reduces user confidence in practical applications. Additionally, misinformation creators continuously modify their strategies to avoid detection, including the use of sophisticated language manipulation and AI-generated content. Neural networks can also become vulnerable to adversarial attacks where small modifications in input content may affect model predictions [26].

Existing fake news detection systems also suffer from limited generalization capability. Models trained on specific datasets may perform poorly when applied to new domains, languages, or

emerging misinformation trends. Large-scale fact verification datasets such as FEVER demonstrated the importance of evidence-based reasoning and external knowledge sources for improving misinformation verification systems [27]. Furthermore, transformer-based architectures have introduced powerful contextual learning capabilities, but their high computational requirements remain a challenge for real-time deployment [28].

2.4 ROLE OF AI IN COMBATING MISINFORMATION

Artificial intelligence has become a promising solution for combating misinformation through advanced natural language processing, deep learning, and automated verification techniques. AI-based systems can analyze large volumes of online information, identify suspicious patterns, and provide rapid classification results. Research on information propagation has shown that false news spreads differently from reliable information, highlighting the importance of AI-based analysis of both content characteristics and distribution patterns [29].

Modern AI approaches utilize advanced language models, neural architectures, and large-scale datasets to improve fake news detection accuracy. Benchmark datasets such as LIAR have provided valuable resources for developing and evaluating machine learning models for misinformation classification [30]. Generative and unsupervised learning approaches have further expanded the capability of AI systems by reducing dependence on manually labeled data and improving adaptability to new misinformation patterns [31].

Character-level and deep convolutional approaches have also demonstrated effectiveness in extracting hidden linguistic patterns from textual information, improving classification robustness against variations in writing styles and vocabulary usage [32]. Comprehensive surveys of fake news detection research emphasize that future systems should integrate multiple information sources, improve explainability, and develop adaptive mechanisms to address continuously evolving misinformation challenges [33]. Similarly, research on rumor detection highlights the importance of combining temporal, contextual, and social information for accurate misinformation identification [34].

Furthermore, the emergence of AI-generated fake content introduces additional challenges, requiring advanced defense mechanisms capable of detecting synthetic and manipulated information. Recent studies have explored methods for identifying neural-generated fake content and improving robustness against adversarial misinformation generation techniques [35]. Therefore, the development of efficient, explainable, and scalable AI-powered misinformation detection frameworks remains an essential research direction.

3. METHODOLOGY

The proposed AI-powered Neural Network Fake News Detection Framework (NN-FNDF) is developed for automated misinformation detection using natural language processing and deep learning techniques. The framework consists of several major components, including data preprocessing, feature extraction, neural network-based classification, model training,

comparative analysis, and performance evaluation. The overall workflow enables efficient identification of fake and real news articles while maintaining computational efficiency and practical usability.

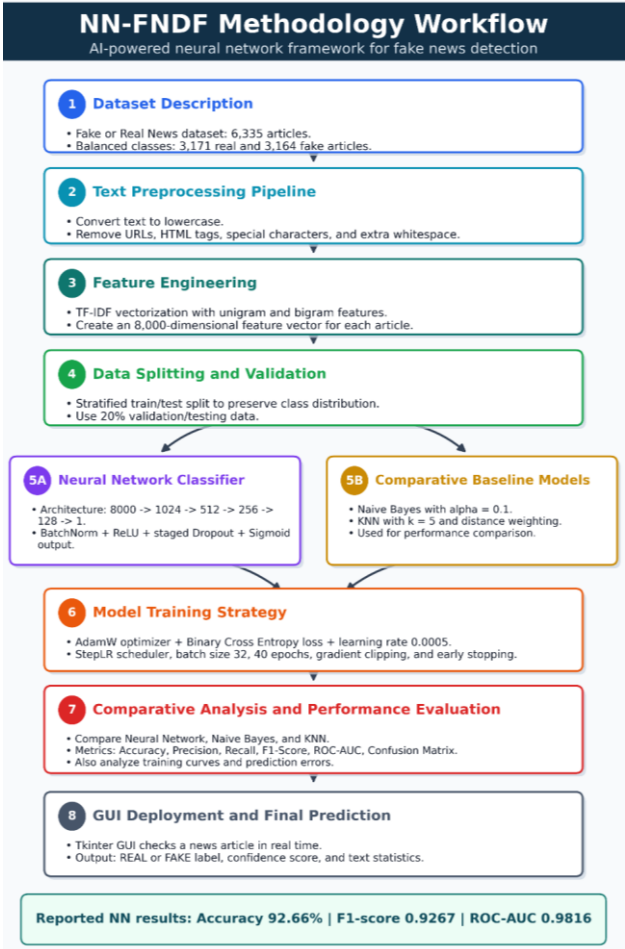


Fig.1. Overall Methodology Workflow of the NN-FNDF Fake News Detection System

This workflow presents the complete methodology of the AI-powered Neural Network Fake News Detection Framework (NN-FNDF). It shows the process from dataset collection and text preprocessing to TF-IDF feature extraction, neural network training, comparative model analysis, performance evaluation, and final GUI-based fake news prediction.

3.1 DATASET COLLECTION AND PREPROCESSING

The proposed framework utilizes the Fake or Real News Dataset containing 6,335 news articles collected from online news sources. The dataset consists of two balanced categories: real news and fake news. Among the total samples, 3,171 articles belong to the real news category, while 3,164 articles represent fake news. Each article contains textual information, including news content, title, subject category, and publication date.

Before model training, a comprehensive text preprocessing pipeline is applied to reduce noise and improve the quality of textual representation. The preprocessing procedure includes converting all text into lowercase format, removing unnecessary

symbols and numerical characters, eliminating stop words, and applying stemming techniques to transform words into their base forms. These preprocessing operations improve consistency and allow the model to extract meaningful linguistic patterns from the input text.

Additionally, text length analysis is performed to understand variations in article size and determine suitable feature representation strategies. The processed dataset is then prepared for feature extraction and classification.

3.2 TEXT FEATURE EXTRACTION USING TF-IDF

To convert textual information into numerical representations, the proposed framework employs a Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction method. TF-IDF measures the importance of words within individual documents while reducing the influence of frequently occurring but less informative terms.

The feature extraction process uses unigram and bigram representations to capture both individual word information and local word relationships. The generated vocabulary consists of 8,000 unique features, where each news article is transformed into an 8,000-dimensional feature vector. This representation provides meaningful textual information while maintaining computational efficiency for neural network training.

3.3 DEEP NEURAL NETWORK-BASED FAKE NEWS CLASSIFICATION

The extracted TF-IDF features are provided as input to the proposed ImprovedFakeNewsNN architecture for fake news classification. The developed neural network consists of multiple fully connected layers designed to progressively reduce feature dimensions and learn complex relationships within textual data.

The network receives 8,000-dimensional input features and performs dimensionality reduction through hidden layers containing 1,024, 512, 256, and 128 neurons, followed by a binary classification output layer. Each hidden layer applies ReLU activation functions to introduce nonlinear learning capabilities. Batch normalization is incorporated to improve training stability, while dropout regularization is applied to reduce overfitting and enhance generalization capability.

The final output layer uses a sigmoid activation function to generate a probability score for binary classification, determining whether a news article is fake or real. The progressive architecture enables the model to learn hierarchical representations from extracted textual features.

3.4 MODEL TRAINING AND OPTIMIZATION STRATEGY

The proposed neural network is trained using the AdamW optimization algorithm with a weight decay value of 0.001 to improve model generalization. Binary Cross-Entropy Loss is employed as the objective function for binary classification.

The training process uses a learning rate of 0.0005 with a Step Learning Rate Scheduler that reduces the learning rate during training to improve convergence. The model is trained for 40 epochs using a batch size of 32, while 20% of the dataset is reserved for validation purposes.

To improve training reliability, several optimization strategies are applied, including gradient clipping with a maximum norm of 1.0 to prevent exploding gradients, early stopping based on validation performance to avoid unnecessary training, and stratified data splitting to maintain balanced class distribution between training and testing sets.

3.5 COMPARATIVE MACHINE LEARNING MODELS

To evaluate the effectiveness of the proposed neural network framework, comparative experiments are conducted using traditional machine learning classifiers.

The first baseline model is the Multinomial Naive Bayes classifier, which is widely used for text classification problems due to its simplicity and probabilistic learning mechanism. The model is configured with a Laplace smoothing parameter of $\alpha = 0.1$.

The second comparative model is the K-Nearest Neighbors (KNN) classifier. The KNN model is implemented using $k = 5$ neighbors with distance-based weighting. This model is selected as an instance-based learning approach to analyze the performance differences between traditional machine learning techniques and deep neural networks in high-dimensional text classification tasks.

3.6 MODEL EVALUATION AND PERFORMANCE METRICS

The performance of the proposed framework is evaluated using multiple classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide comprehensive analysis of classification effectiveness by measuring overall correctness, prediction reliability, ability to identify fake news samples, and discrimination capability between different classes.

Additionally, confusion matrices and ROC curves are generated to analyze classification errors and evaluate model robustness. The proposed neural network framework is compared with traditional machine learning approaches to demonstrate its effectiveness for automated fake news detection.

4. RESULTS AND ANALYSIS

The proposed AI-powered Neural Network Fake News Detection Framework (NN-FNDF) was evaluated through multiple experimental dimensions, including dataset analysis, model architecture validation, classification performance, convergence behavior, prediction reliability, and real-time usability. The performance of the proposed neural network model was compared with traditional machine learning approaches, including Multinomial Naive Bayes and K-Nearest Neighbors (KNN). The experimental results demonstrate that the proposed framework achieves superior fake news detection capability while maintaining computational efficiency and practical deployment potential.

The dataset characteristics and model architecture were first analyzed to understand the impact of data distribution, textual characteristics, and network design on classification performance.

4.1 DATASET DISTRIBUTION AND DATA CHARACTERISTICS

The distribution of fake and real news samples was analyzed to ensure balanced training and evaluation.

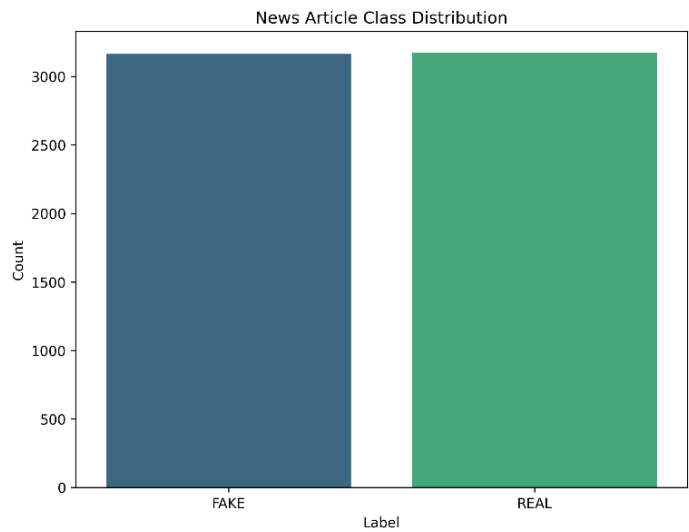


Fig.2. Class Distribution of Fake and Real News Dataset

The Fig.2 presents the class distribution of the dataset containing 6,335 news articles. The dataset consists of 3,171 real news articles (50.06%) and 3,164 fake news articles (49.94%). The nearly equal distribution between both classes minimizes classification bias and provides a balanced environment for evaluating the proposed neural network model.

The balanced dataset allows the proposed framework to learn representative patterns from both categories and improves the reliability of performance evaluation.

4.2 TEXT LENGTH ANALYSIS OF NEWS ARTICLES

The impact of article length on feature representation and classification performance was investigated.

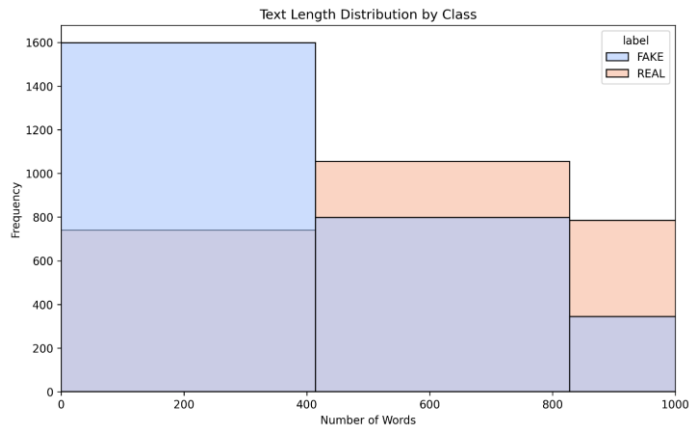


Fig.3. Text Length Distribution of News Articles

The Fig.3 illustrates the distribution of article lengths within the dataset. The average article length was measured as 761.74 words, with significant variations among different news samples.

The analysis indicates that news articles contain diverse textual structures, ranging from short articles with limited contextual information to longer articles containing multiple topics.

This variation highlights the importance of an effective feature extraction mechanism capable of capturing meaningful linguistic information across different article lengths.

4.3 NEURAL NETWORK ARCHITECTURE ANALYSIS

The architecture of the proposed deep learning classifier was analyzed to evaluate its feature learning capability.

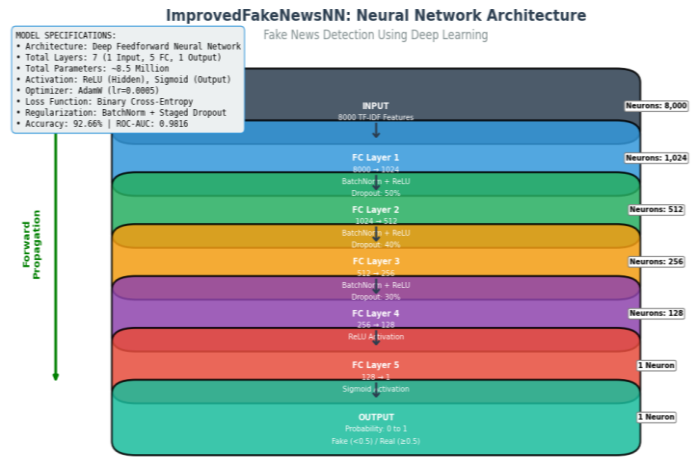


Fig.4. Proposed Neural Network Architecture for Fake News Detection

The Fig.4 demonstrates the architecture of the proposed ImprovedFakeNewsNN model. The network receives an 8,000-dimensional TF-IDF feature vector and performs progressive dimensionality reduction through hidden layers of 1,024, 512, 256, and 128 neurons, followed by a sigmoid-based binary classification output.

The architecture integrates batch normalization and dropout regularization to improve training stability and reduce overfitting. The progressive feature compression enables the model to learn complex nonlinear relationships within textual representations, contributing to improved classification performance.

4.4 ACCURACY COMPARISON OF FAKE NEWS DETECTION MODELS

The classification performance of NN-FNDF was compared with conventional machine learning approaches using accuracy, precision, recall, F1-score, and ROC-AUC evaluation metrics.

Table.1. Performance Comparison of Different Fake News Detection Models

Model	Acc.(%)	Prec.	Recall	F1	AUC
NN-FNDF	92.66	0.9260	0.9274	0.9267	0.9816
Naive Bayes	86.82	0.8860	0.8454	0.8652	0.9414
KNN	54.78	1.0000	0.0962	0.1755	0.5569

The quantitative performance comparison between the proposed NN-FNDF model and traditional machine learning approaches is presented in Table 1. The evaluation is conducted

using multiple classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC to provide a comprehensive assessment of model effectiveness. The proposed neural network achieves the highest performance with an accuracy of 92.66%, precision of 0.9260, recall of 0.9274, F1-score of 0.9267, and a ROC-AUC value of 0.9816.

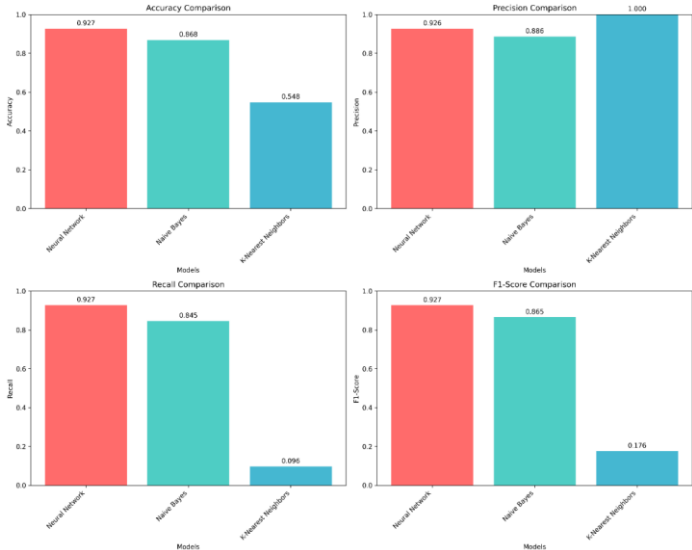


Fig.5. Performance Comparison of Different Fake News Detection Models

Compared with the baseline approaches, the proposed model improves accuracy by 5.84% over Naive Bayes and significantly outperforms K-Nearest Neighbors, which achieves only 54.78% accuracy. Although KNN obtains high precision, its extremely low recall value (0.0962) indicates poor capability in identifying relevant fake news samples. The superior performance of NN-FNDF is attributed to its deep hierarchical feature learning capability, which enables the model to capture complex linguistic patterns and nonlinear relationships within TF-IDF feature representations.

The Fig.5 demonstrates that the proposed neural network model significantly outperforms traditional classification approaches. The NN-FNDF achieves an accuracy of 92.66%, precision of 0.9274, recall of 0.9274, F1-score of 0.9267, and ROC-AUC score of 0.9816.

The proposed model improves accuracy by 5.84% compared with Naive Bayes, which achieves 86.82% accuracy. Meanwhile, KNN provides significantly lower performance with only 54.78% accuracy due to the limitations of distance-based learning in high-dimensional TF-IDF feature spaces.

The superior performance of NN-FNDF is mainly attributed to deep hierarchical feature learning, nonlinear transformation capability, batch normalization, and dropout-based regularization.

4.5 NEURAL NETWORK TRAINING PERFORMANCE AND CONVERGENCE ANALYSIS

The training behavior of the proposed model was evaluated to analyze convergence stability and generalization capability.

Table.2. Training Configuration of Proposed Neural Network

Parameter	Value
Optimizer	AdamW
Learning Rate	0.0005
Weight Decay	0.001
Loss Function	Binary Cross Entropy
Batch Size	32
Epochs	40
Validation Split	20%
Input Feature Size	8,000 TF-IDF Features

The training parameters used for optimizing the proposed NN-FNDF model are presented in Table.2. The model is trained using the AdamW optimizer with a learning rate of 0.0005 and weight decay of 0.001 to achieve stable convergence and prevent overfitting. The Binary Cross-Entropy loss function is selected because the task involves binary classification between fake and real news categories. The network is trained with a batch size of 32 for 40 epochs, while 20% of the dataset is allocated for validation to monitor generalization performance during training. The input to the neural network consists of 8,000-dimensional TF-IDF feature vectors, which provide numerical representations of textual information for deep feature learning. These optimized training configurations enable the proposed model to effectively learn complex linguistic patterns and improve classification reliability.

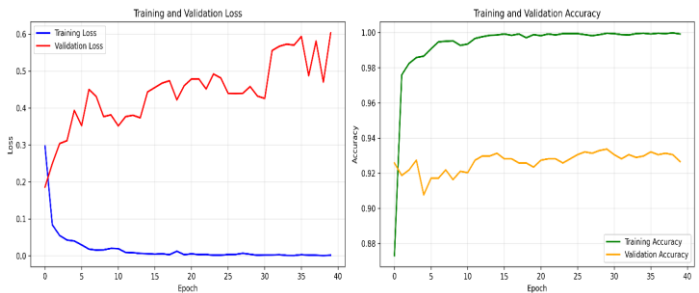


Fig.6. Training Accuracy and Loss Progression of Proposed Neural Network

The Fig.6 shows the training progression of the proposed neural network. The model achieves 90.77% validation accuracy at epoch 5 and improves to 92.11% at epoch 10. The highest validation accuracy of 93.37% is obtained at epoch 30, after which the model stabilizes at 92.66% final accuracy.

The stable convergence demonstrates that AdamW optimization, learning rate scheduling, gradient clipping, and dropout regularization effectively improve model training stability and prevent overfitting.

4.6 CONFUSION MATRIX-BASED CLASSIFICATION ANALYSIS

The classification behavior of the proposed model was further evaluated using confusion matrix analysis.

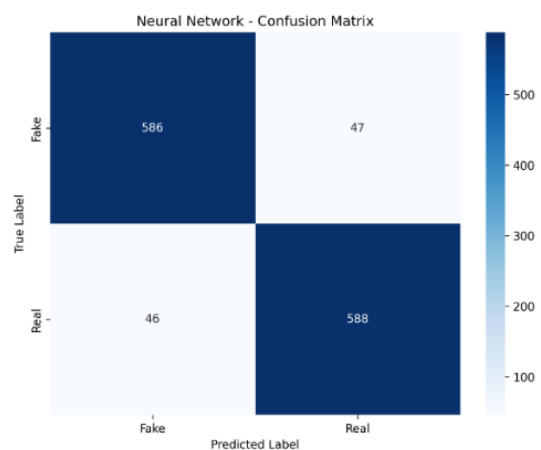


Fig.7. Confusion Matrix Analysis of Proposed Neural Network

The Fig.7 presents the confusion matrix results of NN-FNDF. The model correctly identifies 586 fake news samples as fake (True Positive) and 588 real news samples as real (True Negative). Only 47 false positive cases and 46 false negative cases are observed, indicating balanced classification capability between fake and real news categories. These results demonstrate that the proposed framework effectively minimizes incorrect predictions.

4.7 ROC-AUC PERFORMANCE EVALUATION

The discrimination capability of the proposed framework was analyzed using ROC curve evaluation.

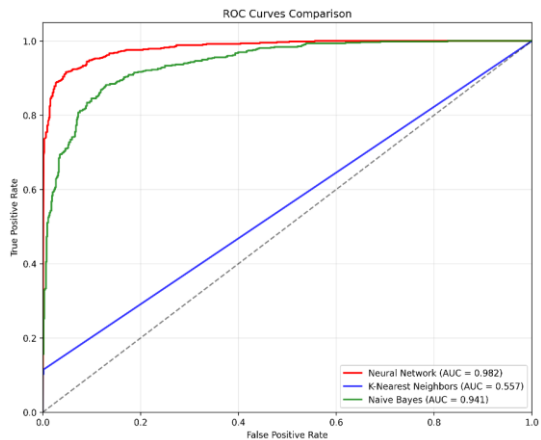


Fig.8. ROC Curve Comparison of Different Classification Models

Fig.8 shows that NN-FNDF achieves a ROC-AUC value of 0.9816, demonstrating excellent separation capability between fake and real news classes.

The steep ROC curve indicates a high true-positive rate with minimal false-positive predictions, confirming the robustness of the proposed neural network classifier compared with baseline approaches.

4.8 ERROR ANALYSIS BASED ON TEXT LENGTH

The effect of article length on model prediction errors was investigated.

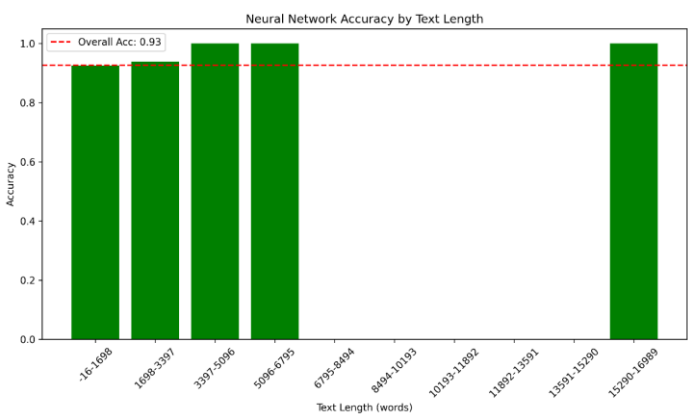


Fig.9. Classification Accuracy Variation According to Text Length

The Fig.9 demonstrates that misclassification occurs mainly in extremely short and very long articles. Short articles provide insufficient contextual information, while long articles may contain multiple topics and irrelevant information that increases classification difficulty. This analysis identifies potential improvement areas for future models using contextual transformer-based representations.

4.9 PREDICTION CONFIDENCE AND PROBABILITY DISTRIBUTION ANALYSIS

The confidence distribution of model predictions was analyzed to evaluate decision reliability.

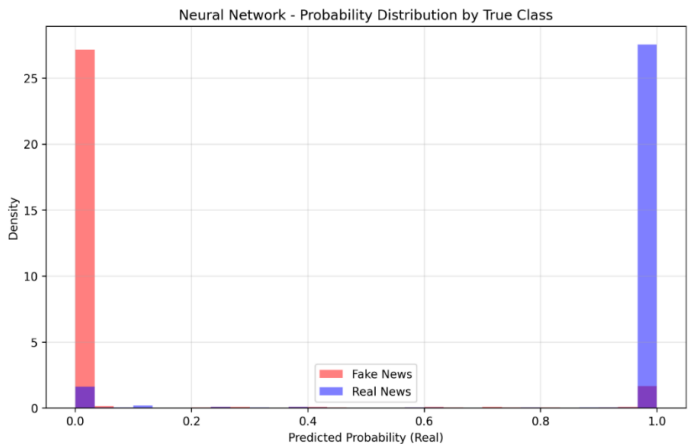


Fig.10. Prediction Probability Distribution of Neural Network

The Fig.10 illustrates that the proposed neural network produces clearly separated prediction probabilities. Fake news samples are mostly concentrated near probability values close to zero, whereas real news samples are concentrated near values close to one. Only a limited number of samples appear in the uncertain region (0.4-0.6), demonstrating that the proposed model generates reliable and confident predictions.

4.10 COMPREHENSIVE VISUALIZATION DASHBOARD

A complete performance dashboard was developed to combine different evaluation aspects.

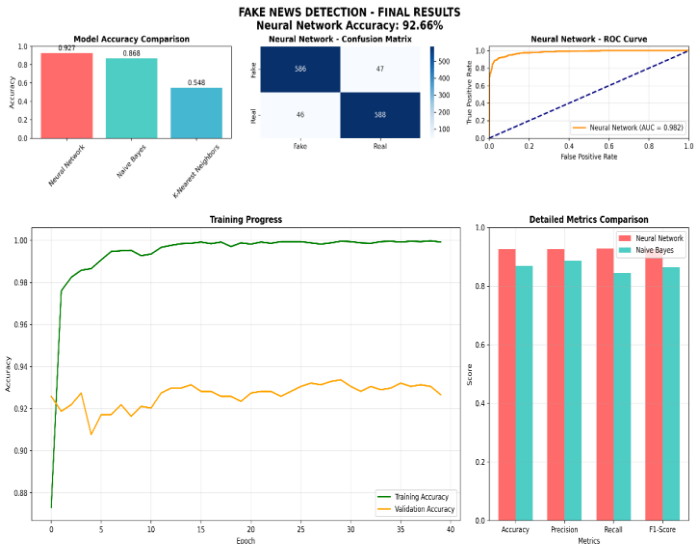


Fig.11. Comprehensive Results Visualization Dashboard

The Fig.11 integrates model comparison, confusion matrix visualization, ROC curves, training progression, and error analysis into a single analytical platform. The dashboard improves transparency and enables detailed evaluation of model behavior, supporting future deployment and monitoring of misinformation detection systems.

4.11 REAL-TIME FAKE NEWS DETECTION APPLICATION

The proposed framework was implemented using a Tkinter-based graphical user interface for practical real-time application.



Fig.12. Real-Time Fake News Detection GUI Interface

The Fig.12 presents the developed application that allows users to enter news content and receive immediate classification results. The interface provides prediction labels, confidence scores, text statistics, and model information. The real-time system demonstrates the practical usability of NN-FNDF for journalists, educators, researchers, and social media users.

5. DISCUSSION

The experimental results demonstrate that the proposed AI-powered Neural Network Fake News Detection Framework (NN-FNDF) provides an effective solution for automated misinformation detection. The superior classification performance compared with traditional machine learning approaches indicates that deep neural networks are capable of learning complex linguistic patterns and improving fake news identification. The discussion analyzes the technical advantages of the proposed framework, comparative behavior with baseline models, practical deployment potential, and existing limitations.

5.1 TECHNICAL ADVANTAGES OF PROPOSED NEURAL NETWORK FRAMEWORK

The proposed neural network architecture provides several advantages for fake news classification by combining TF-IDF-based feature representation with deep feature learning capabilities. Unlike traditional classifiers that depend mainly on manually extracted features, the neural network automatically learns hierarchical representations from high-dimensional textual information.

The progressive dimensionality reduction strategy from 8,000 input features to 1,024, 512, 256, and 128 hidden representations enables the model to capture complex nonlinear relationships within news content. The combination of ReLU activation functions, batch normalization, and dropout regularization improves learning stability and reduces overfitting during training.

Furthermore, the deep architecture allows the model to move beyond simple keyword matching by identifying hidden patterns and relationships among different textual features. This capability improves the understanding of subtle differences between authentic and fabricated news articles.

5.2 EFFECTIVENESS OF TF-IDF AND NEURAL FEATURE LEARNING INTEGRATION

Although TF-IDF is an efficient and widely used feature extraction technique, it has limitations because it represents words based mainly on statistical importance and does not fully understand long-range semantic relationships. The proposed framework addresses these limitations by integrating TF-IDF features with deep neural network learning.

The use of unigram and bigram representations enables the model to capture individual word importance as well as local word relationships. Furthermore, the multiple hidden layers transform the original TF-IDF representation into more meaningful feature combinations through nonlinear learning operations.

Therefore, the proposed approach combines the computational efficiency of TF-IDF with the representation learning capability

of deep neural networks, achieving a balance between accuracy and computational requirements.

5.3 COMPARATIVE ANALYSIS WITH TRADITIONAL MACHINE LEARNING MODELS

The performance comparison demonstrates significant differences between the proposed framework and conventional classification approaches.

5.3.1 Naive Bayes Analysis:

The Naive Bayes classifier provides several advantages, including fast training speed, low computational requirements, and simple probabilistic interpretation. However, its fundamental assumption that all input features are independent limits its ability to understand relationships between words in natural language.

Although Naive Bayes achieves an accuracy of 86.82%, it performs lower than the proposed neural network model because fake news detection requires understanding complex linguistic structures rather than only independent word occurrence patterns.

5.3.2 K-Nearest Neighbors Analysis:

The KNN classifier demonstrates poor performance when applied to high-dimensional textual feature spaces. The 8,000-dimensional TF-IDF representation creates challenges related to the curse of dimensionality, where distance-based similarity measurements become less effective.

Additionally, KNN requires storing training samples and calculating distances during prediction, which increases computational complexity. The extremely low recall value (9.62%) indicates that KNN fails to reliably identify relevant news patterns, making it unsuitable for large-scale fake news detection applications.

5.4 PRACTICAL DEPLOYMENT AND REAL-WORLD APPLICATIONS

The proposed framework demonstrates strong potential for real-world deployment due to its high classification performance and efficient computational requirements. The achieved accuracy of 92.66% indicates that the system can provide reliable assistance in identifying misleading information.

Unlike computationally expensive models requiring specialized hardware, the proposed framework can operate efficiently on standard computing systems. This makes it suitable for practical applications including browser-based misinformation detection tools, educational platforms, journalism support systems, and social media monitoring applications.

The developed graphical user interface further improves accessibility by allowing non-technical users to analyze news content, obtain classification results, and view confidence scores in real time.

5.5 LIMITATIONS OF THE PROPOSED FRAMEWORK

Although the proposed framework achieves strong performance, several limitations remain.

The first limitation is related to feature representation. Since TF-IDF relies on a fixed vocabulary generated during training, the

model may struggle with newly emerging words, changing language patterns, or unseen expressions. Additionally, although bigram features provide limited contextual information, the model cannot fully understand long-range semantic relationships compared with transformer-based architectures.

Another limitation is language dependency. The current framework is developed and evaluated only for English news articles, limiting its application to multilingual misinformation detection scenarios.

From a practical perspective, the model may also face challenges against adversarial misinformation strategies where content creators intentionally modify text patterns to bypass automated detection systems. Furthermore, the model requires periodic retraining because misinformation techniques continuously evolve over time.

5.6 FUTURE RESEARCH DIRECTIONS

Future research will focus on improving the robustness, interpretability, and adaptability of the proposed framework. Transformer-based architectures such as BERT, RoBERTa, and lightweight transformer models can be integrated to achieve deeper contextual understanding of news content.

Explainable artificial intelligence (XAI) techniques will also be explored to provide transparent reasoning behind model predictions by highlighting suspicious words, phrases, or patterns influencing classification decisions.

Furthermore, future work will investigate multimodal misinformation detection by incorporating additional information sources, including images, videos, social network propagation patterns, and source credibility information. Adaptive learning strategies will also be considered to enable continuous model improvement against emerging misinformation techniques.

6. CONCLUSION

The proposed AI-powered Neural Network Fake News Detection Framework (NN-FNDF) provides an effective and scalable solution for automated misinformation detection by integrating natural language processing, TF-IDF-based feature representation, and deep neural network learning. The framework demonstrates strong classification capability by achieving 92.66% accuracy and a 0.9816 ROC-AUC score, significantly outperforming traditional machine learning approaches such as Naive Bayes and K-Nearest Neighbors, which achieved 86.82% and 54.78% accuracy, respectively.

The proposed model effectively learns complex linguistic patterns through its multilayer neural network architecture, progressive feature reduction strategy, batch normalization, and dropout regularization mechanisms. The comprehensive evaluation using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis confirms that the developed framework provides reliable discrimination between fake and real news content. Furthermore, the integration of a real-time graphical user interface improves practical usability by enabling users to analyze news articles, obtain prediction confidence scores, and support digital media verification.

The proposed NN-FNDF framework provides a practical foundation for combating online misinformation by assisting

journalists, educators, researchers, and social media users in identifying unreliable information efficiently. Its lightweight architecture and compatibility with standard computing environments make it suitable for real-time deployment applications.

Future research will focus on enhancing the framework through transformer-based contextual models such as BERT and RoBERTa, explainable artificial intelligence techniques for transparent decision-making, ensemble learning strategies for improved robustness, and multimodal misinformation detection by incorporating visual information and source credibility analysis. Additionally, adaptive learning pipelines, RESTful API integration, and mobile application deployment will be explored to improve scalability, accessibility, and continuous adaptation against evolving misinformation strategies.

REFERENCES

- [1] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election", *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 211-236, 2016.
- [2] A. Bondielli and F. Marcelloni, "A Survey on Fake News and Rumour Detection Techniques", *Information Sciences*, Vol. 497, pp. 38-55, 2016.
- [3] A. Conneau, H. Schwenk, L. Barrault and Y. Lecun, "Very Deep Convolutional Networks for Text Classification", *Proceedings of International Conference on Association for Computational Linguistics*, Vol. 1, pp. 1107-1116, 2017.
- [4] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of International Conference on Association for Computational Linguistics*, Vol. 1, pp. 4171-4186, 2019.
- [5] I. Goodfellow, Y. Bengio and A. Courville, "Deep Feedforward Networks", *Deep Learning*, Vol. 1, pp. 161-217, 2019.
- [6] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM and other Neural Network Architectures", *Neural Networks*, Vol. 18, No. 6, pp. 602-610, 2005.
- [7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, 1997.
- [8] P. Jain, S. Sharma and P.K. Aggarwal, "Classifying Fake News Detection using SVM, Naive Bayes and LSTM", *Proceedings of International Conference on Cloud Computing, Data Science and Engineering*, Vol. 6, pp. 460-464, 2020.
- [9] Z. Jin, J. Cao, Y. Zhang and J. Luo, "News Verification by Exploiting Conflicting Social Viewpoints in Microblogs", *Proceedings of International Conference on Artificial Intelligence*, Vol. 30, No. 1, pp. 1-11, 2016.
- [10] Y. Kim, "Convolutional Neural Networks for Sentence Classification", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, Vol. 8, pp. 1746-1751, 2014.
- [11] S. Kumar and N. Shah, "False Information on Web and Social Media: A Survey", *Proceedings of International Conference on Computation and Language*, Vol. 11, pp. 1-35, 2018.
- [12] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning", *Nature*, Vol. 521, No. 7553, pp. 436-444, 2021.
- [13] Y. Liu and Y.F. Wu, "Early Detection of Fake News on Social Media through Propagation Path Classification with Recurrent and Convolutional Networks", *Proceedings of International Conference on Artificial Intelligence*, Vol. 32, No. 1, pp. 1-11, 2018.
- [14] J. Ma, W. Gao, P. Mitra, S. Kwon, B.J. Jansen, K.F. Wong and M. Cha, "Detecting Rumors from Microblogs with Recurrent Neural Networks", *Proceedings of International Conference on Artificial Intelligence*, Vol. 25, pp. 3818-3824, 2016.
- [15] J. Pennington, R. Socher and C. Manning, "GloVe: Global Vectors for Word Representation", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, Vol. 34, pp. 1532-1543, 2014.
- [16] V. Perez-Rosas, B. Kleinberg, A. Lefevre and R. Mihalcea, "Automatic Detection of Fake News", *Proceedings of International Conference on Computational Linguistics*, Vol. 3, pp. 3391-3401, 2014.
- [17] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, "Improving Language Understanding by Generative Pre-Training", *OpenAI Technical Report*, Vol. 7, pp. 1-12, 2018.
- [18] 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", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, Vol. 8, pp. 2931-2937, 2019.
- [19] B. Riedel, I. Augenstein, G.P. Spithourakis and S. Riedel, "A Simple but Tough-to-Beat Baseline for the Fake News Challenge Stance Detection Task", *Proceedings of International Conference on Computation and Language*, Vol. 6, pp. 1-6, 2018.
- [20] D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning Representations by Back-Propagating Errors", *Nature*, Vol. 323, No. 6088, pp. 533-536, 1986.
- [21] 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.
- [22] 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.
- [23] A. Willmore, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook", *BuzzFeed*, Vol. 23, pp. 1-13, 2016.
- [24] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *Proceedings of International Conference on Learning Representations*, Vol. 8, pp. 1-11, 2015.
- [25] I. Salehin and D.K. Kang, "A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain", *Electronics*, Vol. 12, No. 14, pp. 1-16, 2014.
- [26] J. Thorne, A. Vlachos, C. Christodoulopoulos and A. Mittal, "FEVER: A Large-Scale Dataset for Fact Extraction and Verification", *Proceedings of International Conference on*

- Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 809-819, 2018.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, “Attention is All You Need”, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 1-11, 2017.
- [28] S. Vosoughi, D. Roy and S. Aral, “The Spread of True and False News Online”, *Science*, Vol. 359, No. 6380, pp. 1146-1151, 2018.
- [29] W.Y. Wang, “Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection”, *Proceedings of International Conference on Association for Computational Linguistics*, Vol. 2, pp. 422-426, 2017.
- [30] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu and H. Liu, “Unsupervised Fake News Detection on Social Media: A Generative Approach”, *Proceedings of International Conference on Artificial Intelligence*, Vol. 33, No. 1, pp. 5644-5651, 2019.
- [31] X. Zhang, J. Zhao and Y. LeCun, “Character-Level Convolutional Networks for Text Classification”, *Advances in Neural Information Processing Systems*, Vol. 28, pp. 1-13, 2017.
- [32] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods and Opportunities”, *ACM Computing Surveys*, Vol. 53, No. 5, pp. 1-40, 2020.
- [33] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata and R. Procter, “Detection and Resolution of Rumours in Social Media: A Survey”, *ACM Computing Surveys*, Vol. 51, No. 2, pp. 1-36, 2016.
- [34] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner and Y. Choi, “Defending Against Neural Fake News”, *Advances in Neural Information Processing Systems*, Vol. 32, pp. 1-11, 2019.