Intelligent Detection of Misinformation in Digital Media Using NLP-Based Text Classification and Machine Learning

Pooja Jalikatti[1], Shahina Indikar[2], Vaibhavi Renake[3]
[1,3] *Student, Computer Science and Engineering, SECAB Institute of Enggineering & technology, 586101, India*
[2] *Assistant Professor, Computer Science and Engineering, SECAB Institute of Enggineering & technology ,586101, India*

***Abstract*—The rapid growth of digital communication platforms and online news portals has significantly increased the availability and sharing of information across the world. However, this advancement has also created a major challenge in identifying and controlling the spread of fake news. Fake news contains misleading or false information that can influence public opinion, create confusion, and reduce trust in digital media sources. Traditional manual verification methods are time-consuming and inefficient due to the large amount of news content generated every day. Therefore, an automated fake news detection system using Artificial Intelligence and Machine Learning techniques is required to classify news information accurately and efficiently. This project proposes a Fake News Detection System using DistilBERT Transformer and Machine Learning Algorithms to identify whether a given news article or text content is fake or genuine. The system applies Natural Language Processing (NLP) techniques for understanding textual information and extracting meaningful features from news data. Initially, the input dataset containing news text and corresponding labels is collected and processed. The text preprocessing phase removes unwanted information such as URLs, special characters, symbols, and unnecessary content. The cleaned text is then converted into numerical feature representations using the DistilBERT-based Sentence Transformer model. Unlike traditional feature extraction techniques such as Bag of Words and TF-IDF, transformer-based embeddings capture the semantic meaning and contextual relationship between words, which improves the overall prediction performance.**

After feature extraction, multiple machine learning algorithms including Logistic Regression, Support Vector Machine (SVM), and Random Forest Classifier are trained using the generated DistilBERT embeddings. The dataset is divided into training and testing sets to evaluate the performance of each model. The models are analyzed using different evaluation metrics such as accuracy, precision, recall, and F1-score. Based on the comparison results, the best-performing model is automatically selected using the highest F1-score and stored for future prediction. This approach ensures that the system uses the most reliable classifier for detecting fake news content. A user-friendly web application is developed using the Flask framework to provide real-time fake news prediction. Users can enter any news text through the web interface, where the system pre-processes the input, extracts transformer-based,features,and predicts whether the information is fake or genuine using the trained model. The application also provides a confidence score for the prediction result. Additionally, SQLite database integration is implemented to store previous prediction records, allowing users to view detection history. The proposed system improves fake news identification by combining advanced transformer-based language representation with efficient machine learning classification techniques. The integration of DistilBERT improves the ability of the system to understand complex text patterns and provides better classification compared to traditional approaches. This project demonstrates the effectiveness of artificial intelligence in reducing misinformation and supports users in verifying the authenticity of digital news content. In the future, the system can be enhanced by integrating real-time news verification, multilingual fake news detection, and advanced deep learning models for improved accuracy and scalability.

***Index Terms*—DistilBERT, NLP, Text Classification, Semantic Embeddings, Misinformation Analysis, Machine Learning, News Verification**.

I. INTRODUCTION

The rapid expansion of digital communication platforms and online news sources has transformed the way information is created, shared, and consumed. Social media, news websites, and online forums enable instant access to information, allowing users to receive updates from across the world in real time. Despite these advantages, the increasing volume of online content has led to the widespread dissemination of misleading and fabricated information, commonly known as fake news.

Fake news refers to intentionally or unintentionally false information presented as authentic news. The rapid spread of such content can influence public opinion, create social unrest, and negatively impact sectors such as healthcare, politics, finance, and education. Due to the enormous amount of digital content generated daily, manual verification of news articles has become impractical and time-consuming. As a result, there is a growing need for intelligent automated systems capable of identifying unreliable information efficiently.

Recent advancements in Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) have provided effective solutions for text classification tasks. Traditional fake news detection approaches often rely on feature extraction techniques such as Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). Although these methods are computationally efficient, they often fail to capture contextual and semantic relationships within textual data.

To address these limitations, transformer-based language models have gained significant attention. DistilBERT, a lightweight version of BERT, generates meaningful contextual representations while requiring fewer computational resources. By leveraging transformer-generated embeddings, machine learning classifiers can better understand linguistic patterns and semantic information present in news content.

This work presents a fake news detection framework that combines DistilBERT-based feature extraction with machine learning algorithms, including Logistic Regression, Support Vector Machine, and Random Forest. The system processes news articles, converts textual information into semantic embeddings, and classifies content as genuine or fake. Model performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Furthermore, a Flask-based web application is developed to provide real-time prediction and user interaction. The proposed approach demonstrates the effectiveness of integrating transformer-based representations with machine learning techniques for reliable fake news detection and contributes toward reducing misinformation in digital environments.

II. RELATED WORK

The rapid growth of digital media platforms has increased the spread of misinformation, making fake news detection an important research challenge. Traditional approaches for fake news identification can be broadly categorized into two groups: (i) content-based methods, where textual features extracted from news articles are used for classification, and (ii) source-based methods, where user behavior, publisher credibility, and social interactions are analyzed. Content-based methods have gained significant attention because they directly examine the news text and can be applied across different domains.

Earlier fake news detection systems relied on conventional machine learning algorithms such as Naïve Bayes, Decision Trees, Logistic Regression, and Support Vector Machines. These methods commonly employed feature extraction techniques including Bag of Words (BoW), N-grams, and TF-IDF representations. Although these approaches achieved reasonable classification accuracy, they were unable to effectively capture contextual meaning and semantic relationships between words. As a result, their performance decreased when dealing with complex linguistic structures and evolving writing styles.

To address these limitations, researchers introduced deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. These models automatically learned feature representations from textual data and improved classification performance compared to traditional machine learning techniques. However, deep learning models often require large training datasets, extensive computational resources, and longer

training time, which limits their practical deployment in resource-constrained environments.

Recent advances in transformer-based language models have significantly improved natural language understanding tasks. Models such as BERT demonstrated superior performance by learning contextual representations of words and sentences through attention mechanisms. Several studies reported improved fake news classification accuracy using BERT embeddings when compared with traditional NLP approaches. Despite their effectiveness, full-scale transformer models involve high computational complexity and memory requirements.

To overcome these challenges, lightweight transformer architectures such as DistilBERT were introduced. DistilBERT preserves most of BERT's language understanding capability while reducing model size and inference time. Existing studies have shown that DistilBERT-generated embeddings provide rich semantic information that can improve text classification performance. However, many existing approaches rely on a single classification model, making it difficult to identify the most suitable classifier for diverse datasets.

Motivated by these observations, the proposed work combines DistilBERT-based sentence embeddings with multiple machine learning algorithms, including Logistic Regression, Support Vector Machine, and Random Forest Classifier. Unlike existing approaches that depend on a single prediction model, the proposed system evaluates multiple classifiers and automatically selects the best-performing model based on F1-score. Furthermore, a Flask-based web application integrated with an SQLite database is developed to provide real-time fake news prediction and prediction history management. The proposed framework aims to improve classification reliability while maintaining computational efficiency and practical usability.

3.Proposed Work

Traditional fake news detection approaches mainly relied on machine learning algorithms such as Logistic Regression, Naïve Bayes, and Support Vector Machines combined with feature extraction techniques like Bag of Words and TF-IDF. Although these methods provided reasonable classification performance, they were limited in capturing the contextual and semantic meaning of textual content. To overcome these limitations, deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks were introduced, enabling automatic feature learning from news articles. However, these models often require large datasets, significant computational resources, and longer training time. Recent advancements in transformer-based architectures, particularly BERT, have significantly improved text classification by generating context-aware representations of language. Despite their high accuracy, BERT-based models are computationally intensive for real-time applications. DistilBERT was developed as a lightweight alternative that retains most of BERT's language understanding capabilities while reducing computational complexity. Motivated by these advancements, the proposed work utilizes DistilBERT embeddings along with multiple machine learning classifiers, including Logistic Regression, Support Vector Machine, and Random Forest, to improve fake news detection accuracy. Furthermore, the system automatically selects the best-performing classifier based on F1-score and integrates it into a Flask-based web application for real-time prediction and efficient news verification.

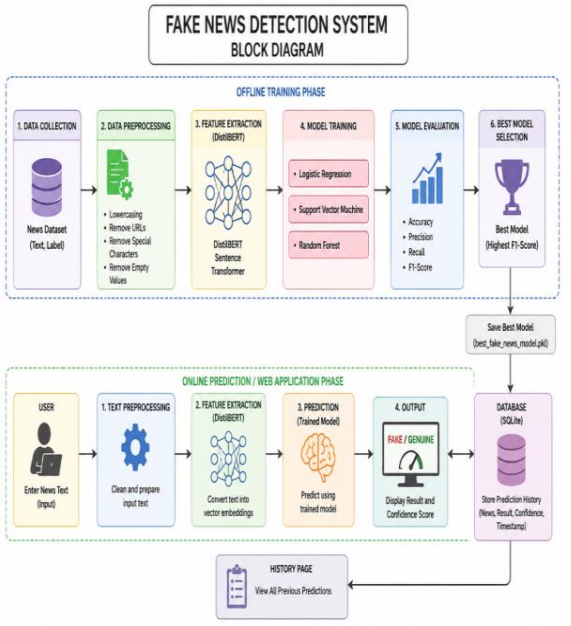

Figure 1: Frame work of proposed work

3.1 Pre-process the dataset

Data preprocessing is an important step used to clean and prepare the collected news text before feature extraction. Raw news data collected from online platforms may contain unwanted characters, symbols, links, missing values, and inconsistent text formats.

These unnecessary elements can reduce the performance of the classification model. In this stage, missing records containing empty news text or labels are removed from the dataset. The news content is converted into lowercase format to maintain uniform text representation. URLs and website links are removed because they do not provide useful information for identifying fake news. Special characters, numbers, and unwanted symbols are eliminated using regular expression techniques. Text cleaning helps remove noise from the dataset and improves the quality of input information. Proper preprocessing allows the model to focus only on meaningful textual patterns. The cleaned news text becomes more structured and suitable for transformer-based feature extraction. This step increases the accuracy and efficiency of the fake news detection system.

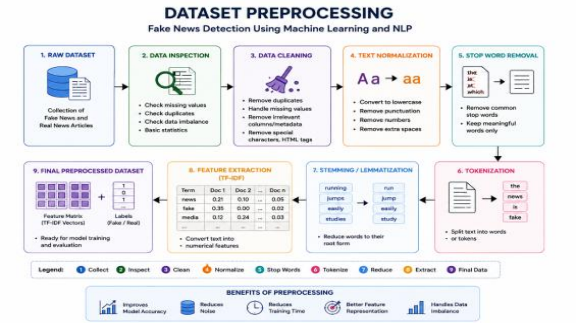

Figure 2: Pre-process the dataset

3.2 Feature Extraction Using DistilBERT Transformer

Feature extraction is the process of converting cleaned textual information into numerical values that can be understood by machine learning algorithms. Since machine learning models cannot directly process human language, text must be transformed into mathematical representations. In this project, the DistilBERT Sentence Transformer model is used for extracting features from news content. DistilBERT is a lightweight transformer model designed to understand the semantic meaning and relationship between words. Unlike traditional methods such as Bag of Words and TF-IDF, DistilBERT analyzes the context of complete sentences instead of only considering word frequency. The cleaned news text is passed into the DistilBERT model, where it generates high-dimensional sentence embeddings. These embeddings represent the meaning and structure of the given news article. Transformer-based features allow the system to understand complex language patterns and improve fake news classification accuracy. The generated feature vectors are provided as input to machine learning algorithms for model training.

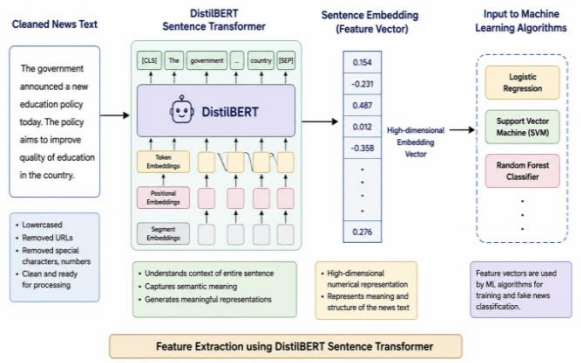

Figure 3: Feature Extraction Using DistilBERT Transformer

3.3 Model Training and Evaluation

Model training is the process of teaching machine learning algorithms to distinguish between fake and genuine news using features extracted from the DistilBERT model. The generated feature vectors are divided into training and testing datasets. The training dataset is used to build three classification models: Logistic Regression, Support Vector Machine (SVM), and Random Forest Classifier. Logistic Regression performs classification by analyzing the relationship between input features and output classes. SVM identifies an optimal decision boundary between fake and genuine news, while Random Forest utilizes multiple decision trees to improve prediction accuracy. Training multiple models helps determine the most effective algorithm for fake news detection.

Model evaluation is conducted to assess the performance of the trained classifiers using unseen testing data. The predictions generated by Logistic Regression, SVM, and Random Forest are compared with actual news labels. Performance is measured using evaluation metrics such as Accuracy, Precision, Recall, and F1-Score. Accuracy indicates the overall correctness of predictions, Precision measures the correctness of positive predictions, Recall evaluates the ability to identify relevant news categories, and F1-Score provides a balanced measure of Precision and Recall. The evaluation results are compared to select the most reliable model for fake news detection.

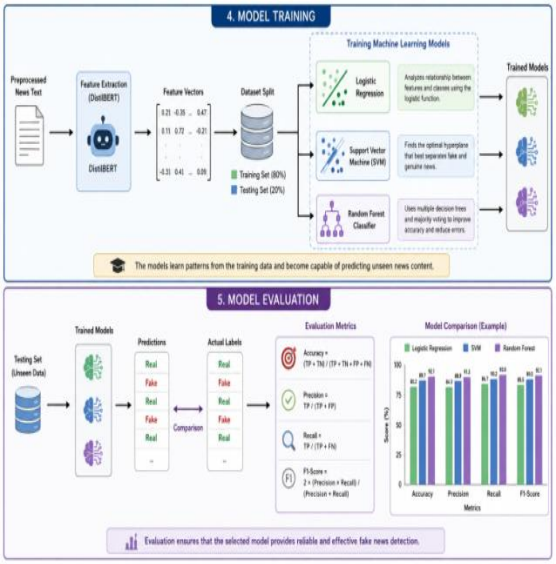

Figure 4: *Model Training and Evaluation*

3.4 Best Model Selection and Storage

After evaluating all machine learning models, the best-performing model is selected based on performance results. In this project, the F1-score value is considered for choosing the best classifier because it provides a balance between precision and recall. The model with the highest F1-score is selected as the final prediction model. Selecting the best model improves the reliability and accuracy of the fake news detection system. After selection, the trained model is stored using the Joblib library. The saved model file allows the system to reuse the trained classifier without performing training again. This reduces processing time and improves application efficiency. The stored model is later integrated with the Flask web application for real-time prediction. Saving the best model provides a reusable and efficient solution for future fake news classification tasks.

3.5. User Input and Web Application

The user interaction phase is developed using the Flask web framework. Flask connects the trained fake news detection model with a user-friendly web interface. Users can enter news text through the webpage for verification. Once the user submits the news content, the application receives the input and sends it for processing. The web interface makes the machine learning system accessible even for non-technical users. The entered news text follows the same preprocessing and feature extraction steps used during model training. Flask handles communication between the frontend interface, trained model, and database. This provides a complete real-time fake news detection environment. The web application improves usability by allowing users to check news authenticity easily.

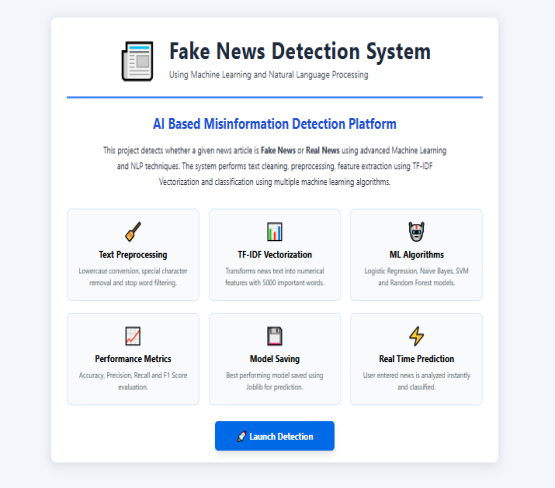

Figure 5. User Input and Web Application

3.6 Prediction and Output Generation

During the prediction phase, the user-entered news text is analyzed by the trained model. Initially, the input text is cleaned using preprocessing techniques. The processed text is converted into numerical embeddings using the saved DistilBERT Sentence Transformer model. These extracted features are passed to the trained machine learning classifier. The classifier analyzes the input pattern and predicts whether the given news is Fake or Genuine. Along with the prediction result, the system calculates a confidence score using prediction probability. The confidence percentage represents how strongly the model supports the generated output. Finally, the result is displayed through the web interface. This process provides users with fast and automated fake news identification.

3.7 SQLite Database and History Management

The final stage of the system involves storing prediction information using the SQLite database. Whenever a user checks any news content, the system stores the entered text, prediction result, and confidence score. SQLite provides lightweight database management without requiring a separate server. The stored information allows users to view their previous fake news detection records. The history feature im-

proves user experience by maintaining past prediction details. Database integration also helps analyze previously checked information. The Flask application communicates with SQLite to insert and retrieve prediction records. This makes the fake news detection system more organized and reliable. The combination of machine learning prediction and database storage provides a complete intelligent web-based solution.

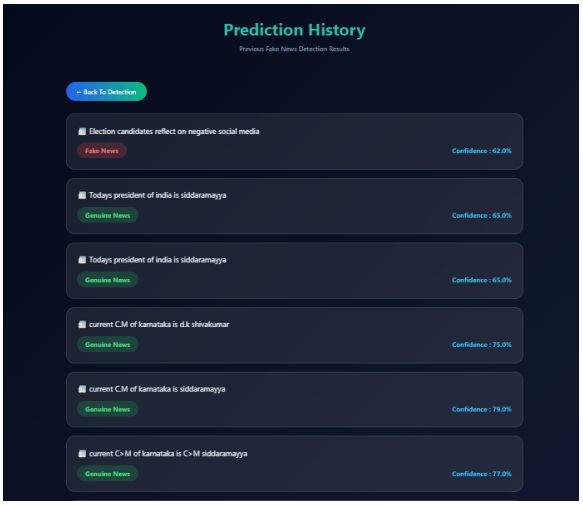

Figure 6:*prediction History Management*

IV. RESULT

The proposed fake news detection system is expected to achieve reliable classification performance by accurately distinguishing fake and genuine news articles. Performance evaluation metrics and comparative analysis will be used to identify the most effective machine learning model. The developed Flask-based application enables users to submit news content and receive classification results along with a confidence score.

The system maintains prediction records using an SQLite database, allowing users to access previously analyzed news items. By integrating Natural Language Processing, DistilBERT-based feature extraction, machine learning algorithms, and web technologies, the platform provides an efficient approach for automated news verification. The proposed framework aims to reduce manual fact-checking efforts and support the identification of misleading information. Future enhancements may include multilingual capabilities, real-time verification mechanisms, and advanced deep learning models to further improve detection accuracy and scalability.

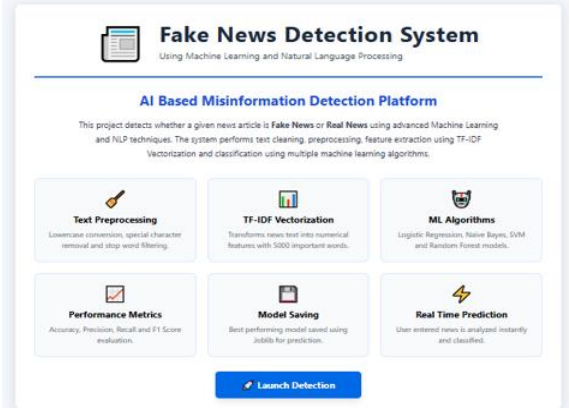

Figure 4.1. Home page

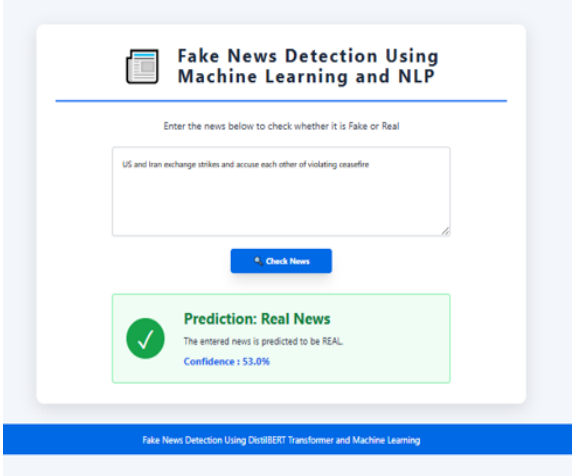

Figure4.2: Real prediction

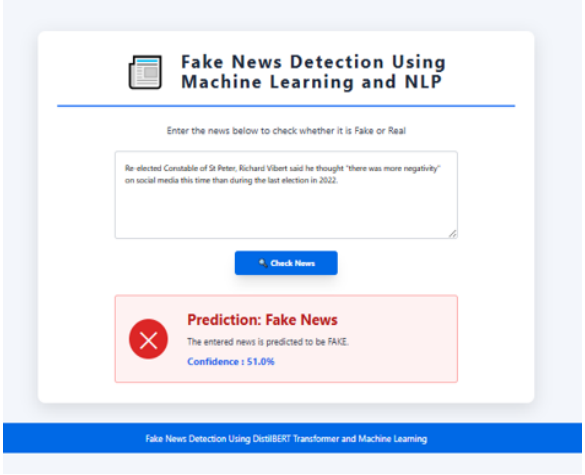

Figure 4.3: Fake Prediction

V. COMPARISON

After evaluating all machine learning models, the best-performing model is selected based on performance results. In this project, the F1-score value is

considered for choosing the best classifier because it provides a balance between precision and recall. The model with the highest F1-score is selected as the final prediction model. Selecting the best model improves the reliability and accuracy of the fake news detection system. Below we have attached the accuracy comparison graph figure11, all metrics comparison graph figure12,f1 score comparison graph figure13, precision comparison graph figure14.

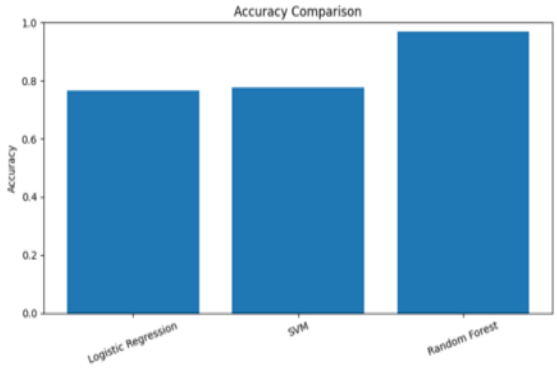

Figure 5.1. Accuracy comparision

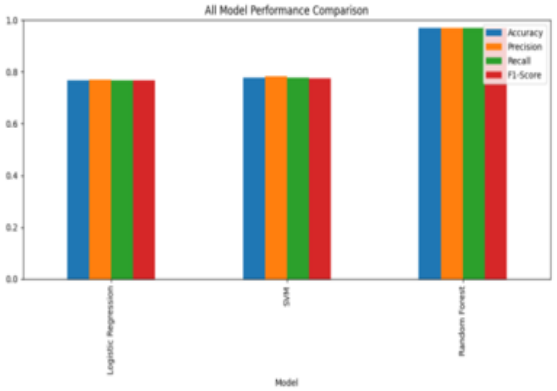

Figure 5.2. All model performance comparison

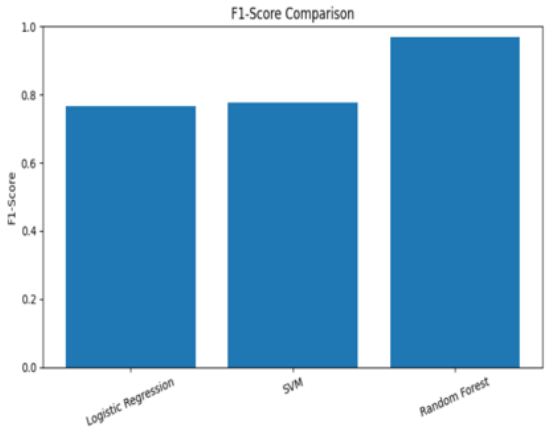

Figure 5.3. F1-score comparision

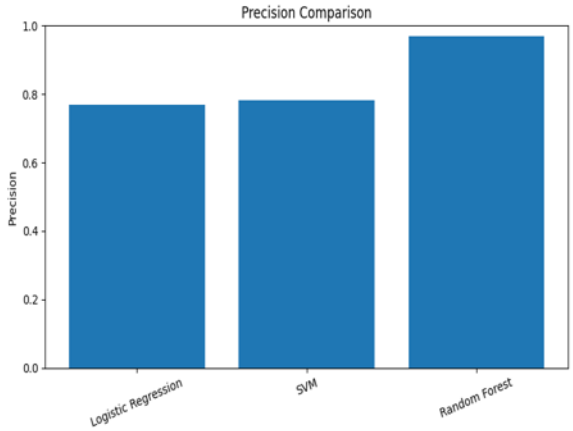

Figure 5.4 .Precision comparision

V. CONCLUSION

The Fake News Detection System using DistilBERT Transformer and Machine Learning Algorithms provides an intelligent and automated approach for identifying misleading and false information from digital news content. With the rapid increase in online communication platforms, the amount of information shared every day has grown significantly, making it difficult for users to manually verify the authenticity of every news article. The spread of fake news creates serious challenges in society by influencing public opinions, creating confusion, and reducing trust in online information sources. This project addresses these challenges by implementing an Artificial Intelligence-based solution that can analyze textual information and classify news as fake or genuine. The proposed system successfully combines advanced Natural Language Processing techniques with Machine Learning algorithms to improve the fake news detection process. Text preprocessing techniques are implemented to clean the raw news data by removing unwanted elements such as URLs, special characters, and unnecessary symbols. This preprocessing stage improves the quality of input data and allows the model to focus on important textual patterns. The cleaned news content is then converted into numerical feature representations using the DistilBERT Sentence Transformer model, which provides a better understanding of semantic meaning and contextual relationships within sentences.

The use of DistilBERT transformer-based embeddings improves the performance of the fake news detection system compared to traditional text repre-

sentation methods. Unlike conventional approaches such as Bag of Words and TF-IDF, DistilBERT captures deeper language information by understanding the relationship between words and their context. This allows the system to identify complex patterns in news articles and make more reliable predictions. The lightweight nature of DistilBERT also makes it suitable for practical implementation with reduced computational requirements. Multiple machine learning algorithms including Logistic Regression, Support Vector Machine, and Random Forest Classifier are implemented and evaluated in this project. Each algorithm is trained using the extracted transformer features and tested to measure its classification performance. The models are compared using evaluation metrics such as accuracy, precision, recall, and F1-score. Selecting the best-performing model based on the highest F1-score ensures that the final system provides efficient and balanced prediction results. The trained model is saved and reused for future predictions, reducing the need for repeated training.

The integration of the trained model with a Flask-based web application makes the fake news detection system accessible and user-friendly. Users can easily enter news content through the web interface and receive instant classification results. The system processes the input text, extracts transformer features, and predicts whether the information is fake or genuine. The addition of a confidence score provides users with a better understanding of the prediction reliability. This real-time detection approach reduces dependency on manual verification methods and improves the speed of identifying false information. The SQLite database integration enhances the functionality of the system by storing previous prediction details such as news content, classification output, and confidence value. This history management feature allows users to review previously analyzed information and improves the overall usability of the application. The combination of machine learning models, transformer-based NLP, database management, and web technologies creates a complete fake news detection platform. In conclusion, the proposed system demonstrates the effectiveness of Artificial Intelligence and Natural Language Processing in solving the problem of misinformation detection. By utilizing DistilBERT for advanced feature extraction and machine learning algorithms for classification, the system provides an efficient, scalable, and reliable solution for detecting fake news. The project contributes to creating a safer digital environment by helping users verify online information before accepting or sharing it. In the future, the system can be further improved by adding real-time news source verification, multilingual support, larger datasets, and advanced deep learning architectures to increase accuracy and handle a wider range of misinformation challenges.

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4171–4186, 2019.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,” *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

[3] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

[4] 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.

[5] H. Ahmed, I. Traore, and S. Saad, “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques,” *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 127–138, 2017.

[6] W. Y. Wang, “Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection,” *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 422–426, 2017.

[7] 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.

[8] M. Granik and V. Mesyura, “Fake News Detection Using Naive Bayes Classifier,” *IEEE First*

Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 900–903, 2017.

[9] Y. Liu and Y. F. B. Wu, "Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks," *AAAI Conference on Artificial Intelligence*, 2018.

[10] S. R. Maulana and A. S. Nugroho, "Fake News Detection Using Machine Learning Approaches: A Systematic Review," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *International Conference on Learning Representations (ICLR)*, 2013.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and others, "Transformers: State-of-the-Art Natural Language Processing," *Proceedings of EMNLP: System Demonstrations*, pp. 38–45, 2020.

[14] A. Conroy, V. Rubin, and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.

[15] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 436–441, 2018.

[16] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," *O'Reilly Media Publications*, 2009.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, and others, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] L. Breiman, "Random Forests," *Machine Learning Journal*, vol. 45, no. 1, pp. 5–32, 2001.

[19] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning Journal*, vol. 20, pp. 273–297, 1995.

[20] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, "Applied Logistic Regression," *John Wiley & Sons Publications*, 2013.