Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

Fake News Detection on Social Media Using Natural Language Processing: A Comparative Study of Machine Learning and Transformer Models

Bellala Prathyusha[1], Dr. Neethu P S[2], Dr. Melbin J Reena[3]

[1, 2, 3] Department of AI and Data Science Engineering, CHRIST (Deemed to be University), Bangalore, Karnataka, India

Abstract:- The heavy reliance on social media for gathering news has enabled quick spread of false information, especially during events such as elections and medical emergencies. Due to large amounts of information on social media, it becomes difficult to test if the information is legitimate or fake. To test the effect of using Natural Language Processing to determine whether a piece of news is true or fake, this study conducted an experiment comparing the efficiency of transformer-based deep learning models with traditional machine learning models. Two benchmark datasets were used in this experiment: the first dataset was FNC-1, for multi-class stance detection, and the second dataset was FakeNewsNet, for binary stance classification. The overall findings indicate that while ensemble machine learning models performed well on the FNC-1 dataset, transformer-based models performed significantly better in determining if the news in the FakeNewsNet dataset is real or fake.

Keywords: DeBERTa, Fake News Detection, Machine Learning, Natural Language Processing, Transformer Models.

1. Introduction

The rapid growth of social media has transformed the production, distribution and consumption of news and information. Social media provides users with fast, easy access to communication and information. It also provides an environment for the rapid spread of false information and fake news. Examples of this kind of misinformation may be seen during the major events in elections, or medical emergencies. Fake news detection refers to the capability of automatically identifying false or misleading news articles among the enormous amount of textual data found in social media formats. The sheer volume of unstructured content available on social networks combined with the prevalence of informal speech patterns, usage of slang, and ambiguity of the context of each post has made this task extremely challenging. Several approaches to fake news detection relied upon traditional machine learning approaches. While these traditional approaches had been successful, they found it challenging to properly identify the deeper semantic and contextual relationships that exist within news articles [8]. Several recent studies [1], [4], [5] have shown that advancements in transformer-based deep learning models have transformed Natural Language Processing by increasing the amount of meaningful representations of textual data. Transformer architectures such as RoBERTa and DeBERTa [5],[6] utilize the contextual embeddings and attention mechanisms. Studies have shown that these models can produce promising results on many different text classification tasks, especially those related to the identification of false news. This paper provides a comparison of two types of models, Transformer model-

Based deep learning and non-deep learning-based machine learning for identifying fake news. This study will determine the best performing methods for each dataset and provide an analysis of their strengths and weaknesses. The main contribution of this work is to find the best models by offering a thorough comparison of Transformer

models with Machine learning techniques under same experimental conditions by methodically assessing and contrasting these methods across various categorization tasks.

Due to growing reliance on social media platforms for the consumption of news, there has been a growing demand for reliable, automated systems for detecting fake news. Because social media platforms do not have strict editorial controls, false information can easily propagate through them and influence people's opinions; something which is not able to happen with traditional media outlets. The spreading of fake news has a serious impact on political decision making, the spread of knowledge regarding public health, and overall stability within the societies. Hence, developing accurate computational techniques for identifying and counteracting fake news has become an area of significant interest and research in the fields of data science and natural language processing. Recent advances in deep learning have produced significant progress in capturing context-specific nuances within text through the latest developments in both supervised and unsupervised learning. Transformer-based architectures have also shown considerable promise as a result of their ability to model both the contextual relationships and the long-range dependencies of text.

Similarly, transformer-based model architectures (e.g., RoBERTa and DeBERTa) exploit self-attention mechanisms for improved contextual understanding to classify news items more accurately than conventional machine learning methods.

2. Related Work

Fake News Detection using Machine Learning (ML) and Deep Learning (DL), have been investigated in detail in numerous published studies in recent years. Earlier studies concentrated on Natural Language Processing (NLP) in combination with traditional machine learning classifiers. According to Naik and Patil [2], text-based feature representation allows for the successful classification of fake news when used together with classifiers such as decision trees, random forests and logistic regression. In addition, Jadhav et al. NLP-based solutions have successfully addressed the challenge of detecting inaccurate content on social platforms, as documented by [8] and others, Studies that utilized ensemble learning have shown that combining predictions from multiple classifiers trained on specific typed 'linguistic features' can lead to increased performance over traditional singular classifier approaches [3], [7], and [11].

Setyadin, et al, [13] demonstrated how machine learning using Random Forest with good Feature Engineering can be competitive regarding the results achieved when using those methods but point out that Altheneyan & Alhadlaq [9] proposed the use of Distributed Learning and Big Data Architecture as solutions to the scalability issues associated with processing large amounts of social media data. In order to overcome the limitations of handcrafted features, researchers started using neural network-based models as deep learning progressed. According to studies by Polu [12] and Prachi et al. [11], word embedding techniques like Word2Vec and GloVe outperform traditional TF-IDF representations in capturing semantic information. The increasing transition from conventional machine learning techniques to deep learning and transformer-based models has been highlighted by recent survey-based investigations that have methodically compiled current fake news detection techniques, datasets, and difficulties [10]. Transformer-based designs have become cutting-edge methods for identifying bogus news in more recent times. While FakeBERT, developed by Kaliyar et al. [14], showed the efficacy of contextualized BERT embeddings for misinformation detection. Kaliyar et al. [1] offered a multimodal transformer- based framework that obtained superior performance across multiple datasets. Further evidence that transformer-based models outperform traditional machine learning techniques by capturing deeper semantic and contextual relationships came from comparative research by Singh and Sharma [5] and benchmark analysis by Shu et al. [6].

Furthermore, recent studies have focused on the model interpretability. Al-Rakhami and Al-Amri [4] proposed an explainable that combines the deep learning and NLP approaches. In order to improve performance and reliability, Raza et al. [15] expanded transformer centric techniques with explainable artificial intelligence mechanisms. Overall, the literature that is now available shows a distinct progression from conventional machine learning techniques to sophisticated transformer-based and explainable models, while highlighting the influence of dataset properties, scalability, and interpretability on the effectiveness of fake news detection.

3. Methodology

This study uses a dual-path experimental architecture to detect fake news using two popular benchmark datasets, FNC-1 and FakeNewsNet. In order to reduce noise and enhance model performance, both methods apply an initial text preprocessing stage that includes stemming, lemmatization, stop-word removal, and lowercasing to clean the data. The first method uses a conventional machine learning pipeline to extract textual characteristics using Hashing Vectorizer representations and Term Frequency–Inverse Document Frequency (TF-IDF). Several classifiers, such as Random Forest, Decision Tree, Logistic Regression, and an Ensemble Voting classifier, are trained using these features. Metrics like accuracy, precision, and F1-score are used to assess model performance. The second method is a transformer-based deep learning framework in which pre-trained transformer models are used to refine the input text after it has been tokenized using the RoBERTa tokenizer. The model is trained and evaluated using the Trainer API from the Hugging Face Transformers library. Both transformer-based models and traditional machine learning techniques are assessed on the same datasets using the same performance indicators to provide a fair comparison. For the purpose of classifying fake news, this experimental setup allows a direct comparison between contextualized transformer representations and feature-based machine learning techniques. The overall dual-path architecture of the proposed system is illustrated in Figure 1.

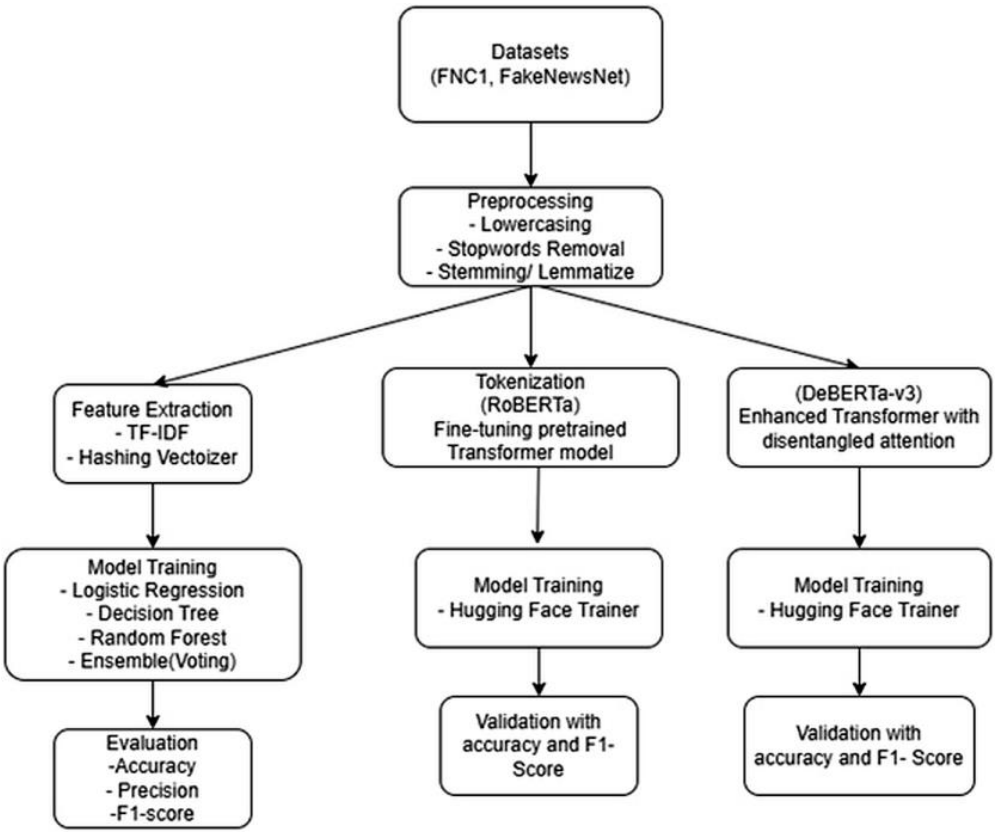

Figure 1. Overall workflow of the proposed fake news detection framework

Feature Extraction:

Feature extraction allows unstructured data to be converted into numeric form for use with classification models. Two classifications of feature extraction techniques were created by this research, as a reflection of the evolution from traditional techniques in Natural Language Processing (NLP) to modern methods built upon Deep Learning techniques. In conventional ma- chine learning (ML) methods, explicit feature engineering will occur alongside common linguistic preprocessing techniques such as lemmatization, stemming, and stop word removal. Once the data has undergone preprocessing steps, various statistical vectorization techniques will be performed on text data to generate sparse numerical representations through the Hashing Vectorizer and Term Frequency Inverse Document Frequency methods.

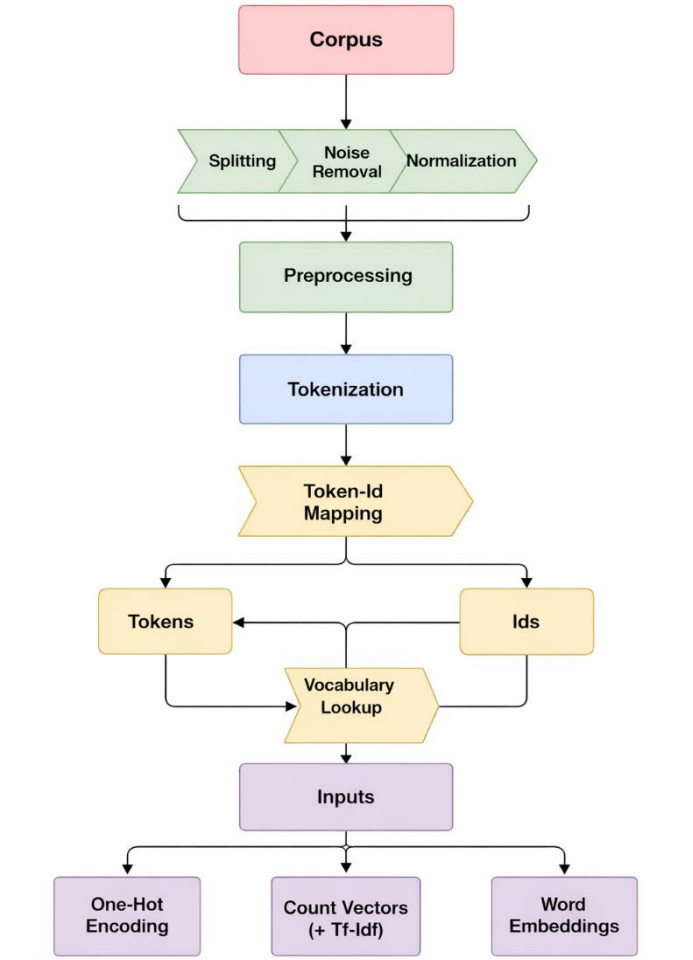

Figure 2. The proposed Fake News Detection framework's Workflow

Hashing Vectorizer:

Hashing Vectorizer is an effective technique in creating fixed-length numerical representation for text documents in machine learning (ML), as well as it being scalable and efficient in terms of memory usage. The Hashing Vectorizer does not maintain an explicit vocabulary within its dictionary; rather, the Hashing Vectorizer calculates a unique index per token (word) based on a hash function. Figure 3 shows how to use the Hashing Vectorizer to transform text into fixed- length feature vectors.

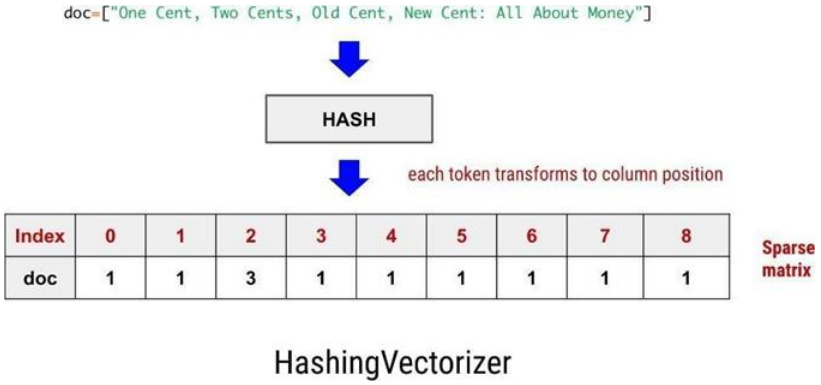

Figure 3. Hashing Vectorizer method, which uses a hash function to transform text tokens

Term Frequency–Inverse Document Frequency:

TF-IDF, a statistical weighting scheme, allows for the identification of the most significant text features in a digital format content Words and phrases that occur with high frequency in a digital document and with less frequency in the corpus, are assigned a higher weight. The presence of a distinct creation of an original Article provides a way to distinguish between fake and authentic news stories. The use of TF- IDF also decreases the effects of frequent use of unhelpful words on the identification of conflicting articles.

Pre-trained Transformer Tokenizers:

Tokenizers based on pre-trained transformers are one of the key elements of the deep learning pipeline to convert unformatted text into valid number tokens that are acceptable for transformer models. Transformers are one example of a neural network architecture that uses a Transformer based approach and contextualised embedding features to represent the input text. Unlike traditional static word embeddings, contextual embeddings are created as a composite of the surrounding words. Each individual word has its own unique representations depending on which context it is used in.

Figure 4 shows the architecture of the transformer- based deep learning model employed in this investigation. By breaking down large, complicated words into meaningful parts, subword tokenization allows transformer models to effectively deal with out-of-vocabulary words. Considering that many pieces of content on social media have unique ways of spelling and abbreviating words (e.g., slang), it is critical that a tokenization method can help build and serve models for training and inference. A tokenizer converts a given piece of text into a set of tokens that can be used to numerically represent the text in a form that can be processed by the model.

Once the textual input has been tokenised, it is converted into embeddings that represent the relationships between the meanings of the words. Regardless of the location of a token within a container (input sequence), transformer architectures apply self attention to observe its dependencies. One of the advantages of this over traditional sequential architectures such as those based on RNNs is that transformers can better understand how to interpret what has been said in terms of context. In addition, positional encoding is used to store the ordering of the words in the input, allowing the model to maintain both the syntax and the meaning of the input.

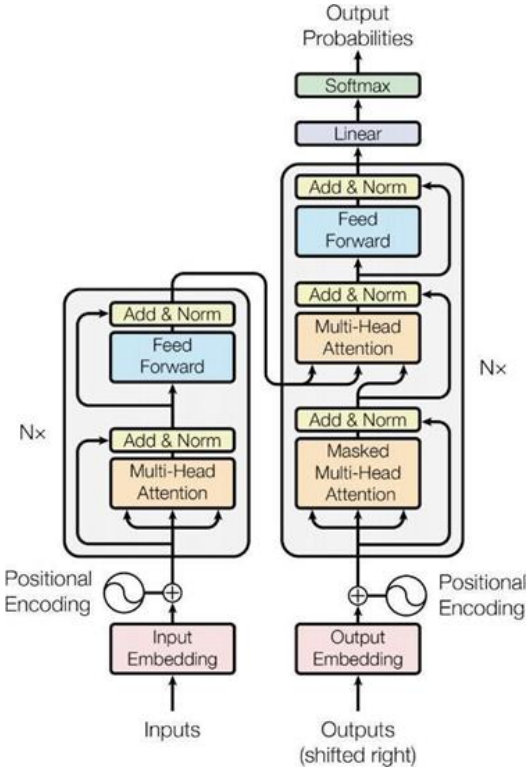

Figure 4. Transformer Model Architecture

Feature Extraction Outcome:

To accommodate both deep learning and traditional machine learning, the proposed system utilizes a dual approach to feature extraction. In traditional ML, explicit feature engineering occurs after extensive preprocessing of the data. The proposed system will clean up textual data so it can be effectively disposed of by using Sparse Statistical Features with Hashing Vectorization and Weighted via TF-IDF. On the contrary, deep learning with transformer-based contextual representations significantly reduces reliance on manual feature engineering. The raw text is processed by the transformer generating sub word level inputs and producing control signals. The transformer model generates contextual embeddings containing the structural, syntactic, and semantic information about the raw text. The improved performance of transformer-based classification systems results from these dense information rich embeddings serving as high quality features.

4. Experimental Results and Discussion

Using two benchmark datasets, FakeNewsNet with binary classification and the Fake News Challenge Stage 1 provided with multi-class stance detection to evaluate the proposed deep learning and machine learning algorithms and a comparative analysis of each algorithm's performance against other algorithms on different datasets, tasks, and features. This provides an opportunity for comparison

between different types of model performances in various task settings and dataset characteristics. Model performance metrics comprise accuracy and F1-score. Given the imbalance across classes, F1-score is an unbiased measure because it combines both precision and recall into one value. For the multi-class FNC-1 dataset, a weighted F1-score was calculated to adjust for the differing class distributions across

all three classes.

FNC-1 Dataset Results:

The FNC-1 dataset, which categorizes news articles into four categories: Agree, Disagree, Discuss, and unrelated were developed to tackle the multi-class stance classification problem. The performance of models trained on the FNC-1 dataset has been evaluated using the weighted F1-Score, which accounts for class imbalance. Ensemble-based machine learning models achieved competitive results (an average of ~92 percent accuracy) when applied to the FNC-1 dataset. Transformer-based model methods have shown the most consistent performance. Out of all of the methods tested for this task, the accuracy of RoBERTa (96.0%) and DeBERTa (96.09%) are both higher than those of the ensemble-based machine learning methods. The easiest to classify across all stance classes by far has been Unrelated, with Agree and Discuss continuing to be the most challenging to predict. Figure 5 shows the accuracy comparison of transformer-based and machine learning models on the FNC-1 dataset. The performance comparison of transformer-based and machine learning models on the FNC-1 dataset is shown in Table 1.

Table 1. Performance comparison of machine learning and transformer-based models on the FNC-1 dataset

Model Type	Model	Accuracy (%)
Machine Learning	Ensemble Classifier	92.0
Transformer	RoBERTa	96.0
Transformer	DeBERTa	96.09

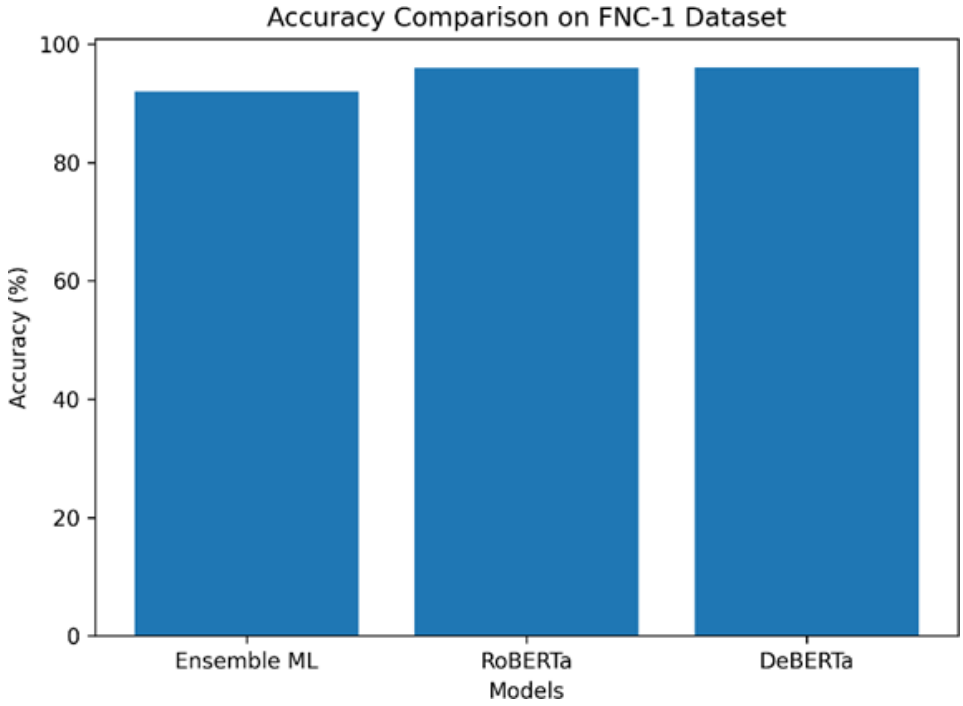

Figure 5. Accuracy comparison of machine learning and transformer-based models on the FNC-1 dataset

FNN Dataset Results:

The FakeNewsNet dataset presents a binary classification problem where news articles must be labeled as either Real or Fake. The challenge in this case is primarily based on language trends and content-related features that are associated with false information. Traditional Machine Learning algorithms using this dataset achieve moderate performance, with approximate accuracy being in the vicinity of 83.0% when an ensemble classifier is used. The use of Transformer-based models exhibit the effectiveness of training contextualised representations for binary false news classification purposes; using the RoBERTa model we achieve 97.0% accuracy and then with the DeBERTa model we get 98.75% accuracy. Figure 6 shows the accuracy comparison of transformer-based and machine learning models on the FakeNewsNet dataset. The Performance comparison of machine learning and transformer-based models on the FakeNewsNet dataset is shown in Table 2.

Table 2. Performance comparison of machine learning and transformer-based models on the FakeNewsNet dataset

Model Type	Model	Accuracy (%)
Machine Learning	Ensemble Classifier	83.0
Transformer	RoBERTa	97.0
Transformer	DeBERTa	98.75

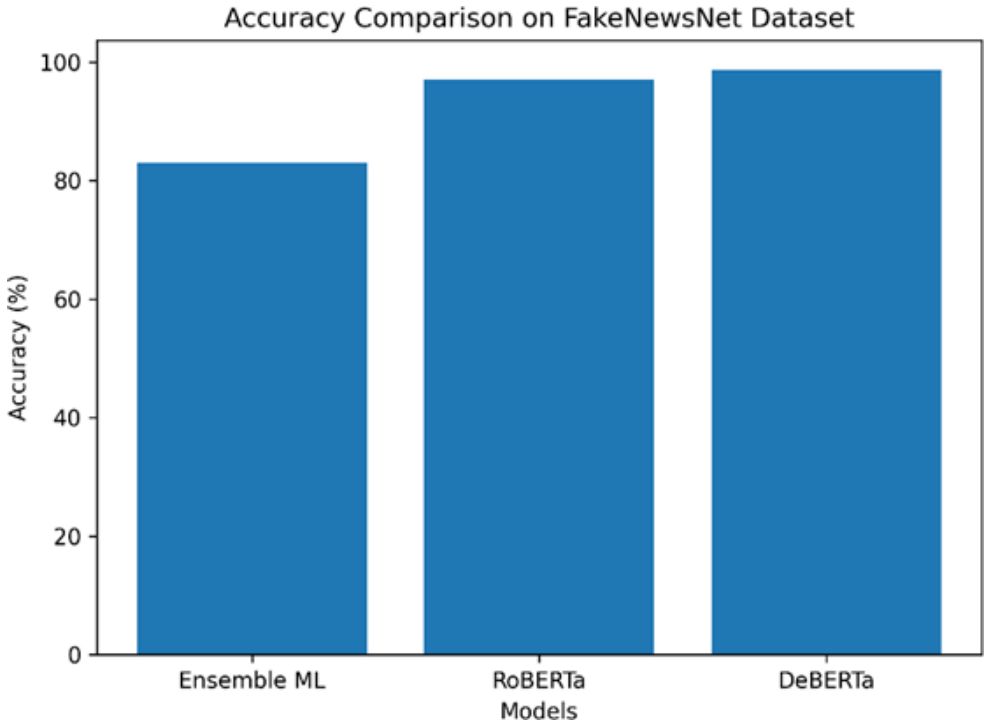

Figure 6. Accuracy comparison of machine learning and transformer-based models on the FakeNewsNet dataset.

5. System and Software Requirements

Traditional (non-transformer) machine learning algorithms were developed using the Scikit-learn package, while transformer-based models were implemented using Pytorch and the Hugging Face API. For visualising the results of the experiments, we utilized the Matplotlib library, as well as other libraries, including NumPy and Pandas, to help pre-process the data prior to training. All experiments were carried out using Python as the programming language. Experiments to evaluate the various models used a computer running Windows 10 and equipped with Intel Core i7 processor.

6. Conclusion

Social Media Fake News Detection Model Development utilizes artificial intelligence to include various modelling approaches, From the results of our investigation, we have observed that transformer-based models outperform classical methods for detecting fake news on FakeNewsNet datasets due to their high level of contextual understanding and interpretation. However, the ensemble machine learning methods demonstrate competitive performance relative to transformer models for detecting fake news on FNC-1 datasets. Out of all the models tested, DeBERTa provided the best overall performance across both the FakeNewsNet and FNC-1 datasets. While transformer models have made significant difference in the field, the available research is limited to only text-based research and a few benchmark datasets which may limit the ability to generalize results to other situations. In conclusion, it can be said that transformer-based models are superior to classical ML techniques for complex cases of detecting fake news, whereas classical ML techniques may work fine in certain structured contexts. It is likely future research will find more datasets, Multimodal Data, and the introduction of Explainable Artificial Intelligence (XAI) techniques may enhance the robustness and interpretable nature of our model.

References

[1] S. Kaliyar, A. Goswami, and P. Narang: Multimodal fake news detection using transformer based architectures. Expert Systems with Applications 196, 116659 (2022).
[2] S. Naik and A.Patil: Fake news detection using NLP In: International Journal of Research in Applied Science and Engineering Technology 9(XII), 8–10 (2021).
[3] B. M. G., R. Harigaran, K. Jeevanantham, V. Sakthivel, P. Sri Varshan, and M. S. Vineeth : Fake news detection using a stacked ensemble of machine learning models. In: Proc. Int. Conf. on Intelligent Data Communication Technologies and Internet of Things (IDCIOT) (2024).
[4] M. Al-Rakhami and A. Al-Amri: Explainable fake news detection using deep learning and NLP techniques Information Processing & Management 60(1), 103144 (2023).
[5] P. Singh and V. Sharma: A comparative study of transformer models for fake news detection in social media. Applied Artificial Intelligence, 38(2), 1-18(2024).
[6] R. Shu, Y. Wang, and H. Liu: A benchmark study of machine learning and transformer- based models for fake news detection. Knowledge-Based Systems 235, 107675 (2022).
[7] Preetham H., Prithviraj T. Chavan, Pranav R., Prathik Vittal, Vikranth B. M.: Fake news detection in social media. International Journal of Engineering Research & Technology (IJERT) 12(06) (June 2023).
[8] Jadhav, T., Adlinge, S., Dande, N., Jadhav, S.: Fake news detection on social media using NLP. International Journal of Scientific Research in Science and Technology 12(6), 175–181 (2025).
[9] A. Altheneyan and A. Alhadlaq: Big data ML-based fake news detection using distributed learning. IEEE Access 11, 32607–32619 (2023).
[10] Shen, Y., Wang, S., Liu, Y.:Fake news detection on social networks: A survey. Applied Sciences 13(21) (2023).
[11] N. N. Prachi, M. Habibullah, M. E. H. Rafi, E. Alam, and R. Khan: Detection of fake news using machine learning and natural language processing algorithms. Journal of Ad vances in Information Technology 13(6), 651–661 (2022).
[12] O. R. Polu: AI-based fake news detection using NLP. International Journal of Artifi cial Intelligence and Machine Learning 3(2), 231–239 (2024).
[13] Setyadin, R.D., Nugroho, A.S., Santoso, B., Wibowo, A.: Fake news detection using Random Forest algorithm. International Journal of Advanced Computer Science and Applications 16(2) (2025).
[14] K. Kaliyar, A. Goswami, and P. Narang: FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications 80, 11765– 11788 (2021).
[15] Raza, N., Singh, A., Gupta, P., Khan, M.: Enhancing fake news detection with trans former-based models. Journal of Artificial Intelligence and Data Science 4(1) (2025).