MAIM - Final Report

Rana Gaber

August 2025

Abstract

This project explores the task of fake news detection by comparing a range of modeling approaches, from classical methods to more advanced architectures. We worked with a Kaggle dataset of real and fake news articles, which was preprocessed by merging titles with article text and applying common text-cleaning steps such as tokenization, lowercasing, and stopword removal. For classical models, Support Vector Machines, Logistic Regression, and Naïve Bayes were trained using two word representations: TF-IDF and word2vec. Sequential models, including RNN, LSTM, and GRU, were then implemented to test their ability to capture temporal dependencies in text. Finally, transformer-based models, namely BERT and DistilBERT, were fine-tuned for the same task. All models were evaluated using accuracy, precision, recall, F1-score, and ROC curves, with additional visualisations such as confusion matrices to better interpret results. The overall study provides a comparative view of how different machine learning paradigms handle the challenges of detecting misinformation.

1 Introduction

The swift dissemination of news and its widespread availability to the public presents a significant ethical dilemma concerning the proliferation of false information, commonly referred to as fake news. This phenomenon underscores the urgent need for the development of tools and methodologies that can effectively differentiate between credible news sources and deceptive content. Such advancements are crucial not only for preserving the integrity of public discourse but also for fostering an informed citizenry capable of making knowledgeable decisions in an increasingly complex media landscape. In light of this, various models and methodologies are scrutinized to assess their efficacy in discerning authentic information from misinformation. In this report, we investigate two distinct word representation methodologies that are integral to the classical model pipeline: Term Frequency-Inverse Document Frequency (TF-IDF) and the skip-gram model derived from word2vec. Furthermore, our experimental framework is broadened to encompass not only traditional models but also sequential architectures and transformer-based models. The classical models evaluated in this report include Support Vector Machines (SVM), Naive Bayes, and Logistic Regression. In contrast, the sequential models consist of Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNN), and Gated Recurrent Units (GRU). Finally, we also examine transformer models that have undergone fine-tuning, specifically BERT and DistilBERT. This comprehensive approach allows for a robust comparative analysis of various machine learning paradigms within the context of natural language processing.

2 Dataset Description

The dataset utilized in this study was sourced from Kaggle, consisting of two distinct files: one representing fake news articles and the other comprising real news articles. The fake news CSV file contained 23502 articles, while the real news file comprised 21417 articles. Each document within these datasets included four key columns: title, text, date, and subject.

2.1 Preprocessed Dataset

The preprocessed dataset underwent a comprehensive series of steps to enhance its suitability for analysis. Specifically, all dates and subjects were excluded from the dataset, while the title was amalgamated with the article text into a singular column. This decision was predicated on the assumption that the title could serve as a valuable contextual element for the model, whereas the inclusion of dates and subjects might introduce noise and mislead the analytical outcomes. Subsequently, the newly combined column underwent several preprocessing procedures, which included: tokenization, case normalization, stop word removal, elimination of numerical values and tags, and, as an optional measure, lemmatization.

In Figure 1, a word cloud is illustrated, effectively representing the frequency of word occurrences, with the size of each word corresponding to its prevalence in the dataset.

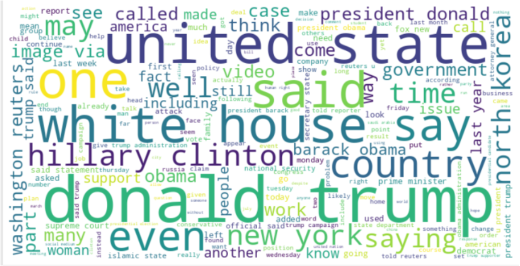


Figure 1: Word cloud

However, in Figure 2, a word frequency plot is presented, which depicts the occurrence rates of the twenty most commonly used words. This visualization highlights the predominant vocabulary within the analyzed text, providing insight into the thematic elements emphasized in the discourse.

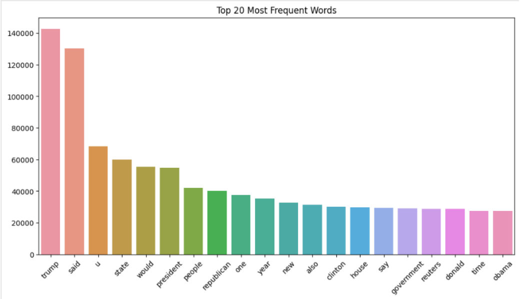


Figure 2: Top-most occurring words in the dataset

3 Methodology

In the current project, we employed three principal methodologies, each characterized by distinct modeling techniques. The approaches included classical models utilizing two different representations of word embeddings, sequential models, and the fine-tuning of transformer architectures. The classical models encompassed support vector machines, logistic regression, and naïve Bayes classifiers, all of which were implemented using the scikit-learn library. Each of these models was evaluated twice, employing each of the aforementioned word representations.

In addition to classical models, we conducted experiments with sequential models to investigate the efficacy of Long Short-Term Memory (LSTM) networks in surpassing the performance of traditional methods, alongside Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs). These sequential models were implemented using TensorFlow.

Finally, we performed fine-tuning on both BERT and DistilBERT architectures, utilizing TensorFlow utilities for this process. The performance of all models was assessed using standard evaluation metrics, including accuracy, F1-score, precision, and recall. Additionally, we presented visual representations of model performance, such as ROC curves and confusion matrices, to facilitate a comprehensive analysis of the results.

4 Results and Error analysis

In this section, we report the performance of each model, highlighting their effectiveness in tackling the problem of fake news detection.

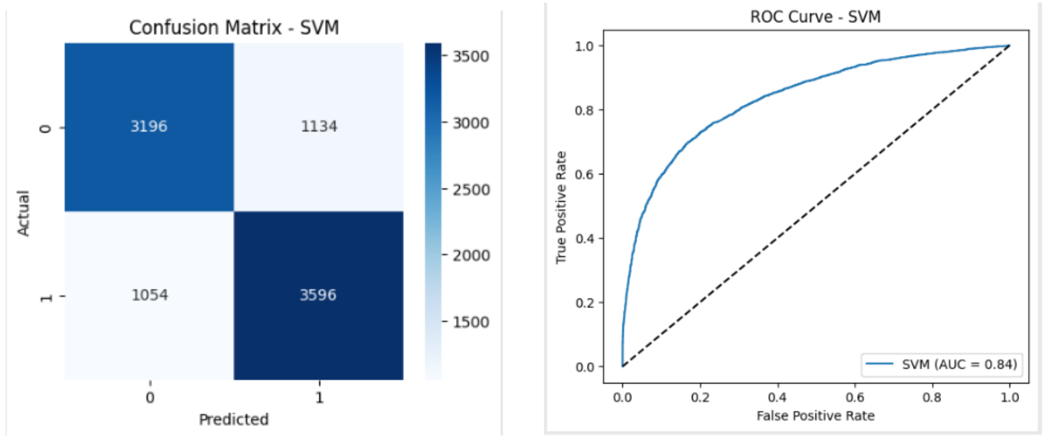
4.1 Classical Models - Word2vec

4.1.1 Support vector machine

The Support Vector Machine (SVM) demonstrated suboptimal performance when applied with the skip-gram word representation. As illustrated in Table 11, the model achieved an accuracy of 75 percent, which is comparatively lower than the metrics reported by alternative models. However, an examination of the Receiver Operating Characteristic curve depicted in Figure 3 reveals that, despite its underwhelming performance, the model's predictions were not arbitrary; the Area Under the Curve exceeded 0.5, indicating some degree of discriminative capability.

Model	Accuracy	Precision	Recall	F1-Score
SVM	75.63	75.63	75.63	75.63

Table 1: Performance of SVM on fake news detection with word2vec word representation



(a) Confusion Matrix (SVM)

(b) ROC Curve (SVM)

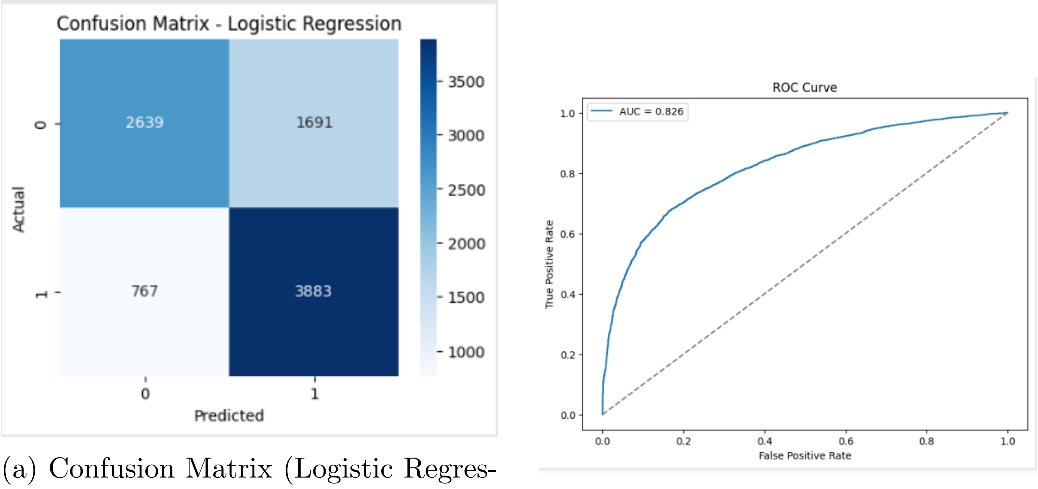
Figure 3: Performance evaluation of the SVM model using confusion matrix and ROC curve.

4.1.2 Logistic Regression

The performance of this Logistic Regression demonstrated a notable decline, as evidenced by a further reduction in all evaluation metrics. However, the Area Under the Receiver Operating Characteristic Curve remained significantly above 0.5, indicating that the model's predictive performance was not merely random. Additionally, the confusion matrix reveals that the ratio of false positives to false negatives approached 0.5, suggesting a lack of substantial bias in the model's predictions.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	72.63	73.43	72.63	72.23

Table 2: Performance of Logistic Regression on fake news detection with word2vec word representation



(a) Confusion Matrix (Logistic Regression)

(b) ROC Curve (Logistic Regression)

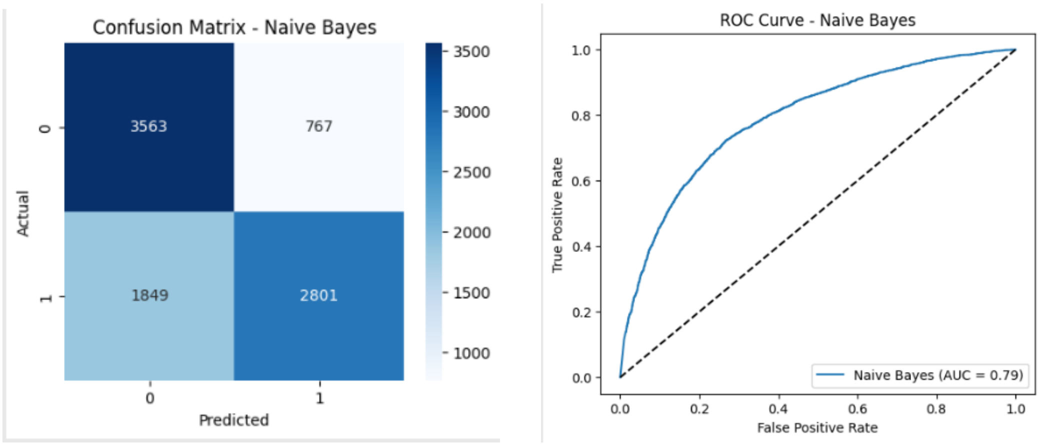
Figure 4: Performance evaluation of Logistic Regression using confusion matrix and ROC curve.

4.1.3 Naive Bayes

Naive Bayes has consistently demonstrated limited efficacy in the context of this task, ranking as one of the two least effective models overall and the weakest among all traditional methodologies. Despite not being a purely random model, comprehensive evaluation metrics indicated subpar performance across the board. Notably, the incidence of false positives was more than twice that of false negatives, suggesting significant concerns regarding the model's reliability and overall effectiveness.

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	70.78	72.39	70.87	70.57

Table 3: Performance of Naive Bayes on fake news detection with word2vec word representation



(a) Confusion Matrix (Naive Bayes)

(b) ROC Curve (Naive Bayes)

Figure 5: Performance evaluation of Naive Bayes using confusion matrix and ROC curve.

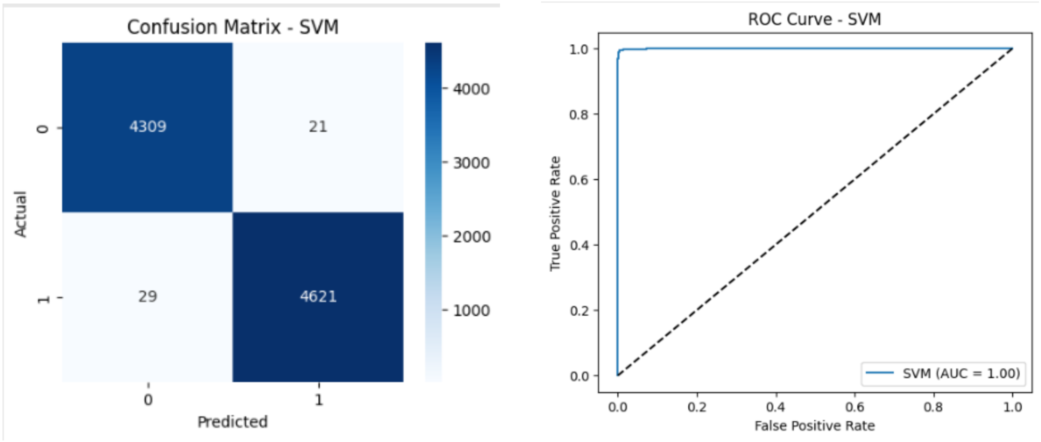
4.2 Classical Models - TF-IDF

4.2.1 Support vector machine

The integration of traditional models alongside TF-IDF representations, as opposed to utilizing word2vec embeddings, yielded significant improvements in performance. Specifically, when employing Support Vector Machines, the evaluation metrics displayed near-perfect results, characterized by an Area Under the Curve being 1. Additionally, the occurrence of misclassified instances was remarkably low, underscoring the effectiveness of this approach in enhancing classification accuracy.

Model	Accuracy	Precision	Recall	F1-Score
SVM	99.44	99.44	99.44	99.44

Table 4: Performance of SVM on fake news detection with TF-IDF word representation



(a) Confusion Matrix (SVM)

(b) ROC Curve (SVM)

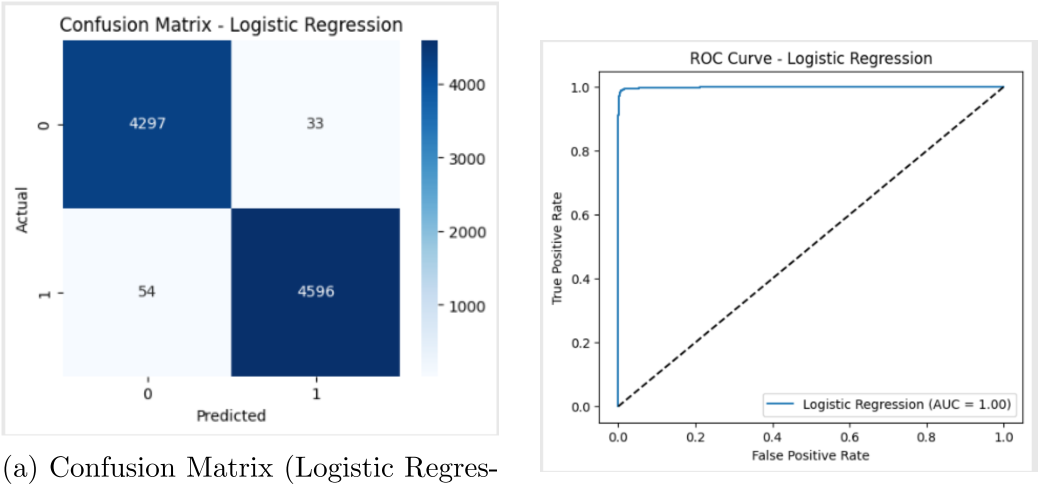
Figure 6: Performance evaluation of SVM using confusion matrix and ROC curve.

4.2.2 Logistic Regression

Although the scores obtained from logistic regression were marginally lower than those achieved by support vector machines (SVM), the evaluation metrics indicated near-perfect accuracy. There were slightly more misclassified samples in the logistic regression model; however, the area under the curve (AUC) remained 1, reflecting a strong predictive performance.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	99.03	99.03	99.03	99.03

Table 5: Performance of Logistic Regression on fake news detection with TF-IDF word representation



(a) Confusion Matrix (Logistic Regression)

(b) ROC Curve (Logistic Regression)

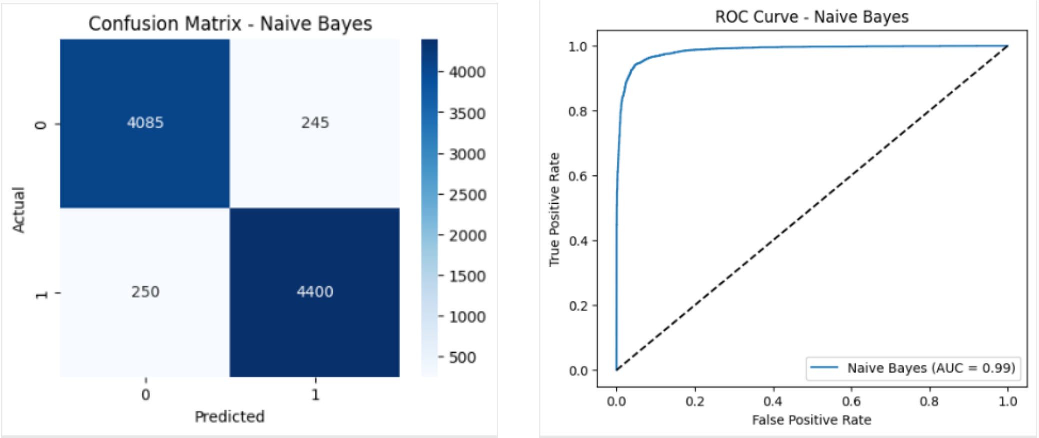
Figure 7: Performance evaluation of Logistic Regression using confusion matrix and ROC curve.

4.2.3 Naive Bayes

The performance metrics obtained from the Naive Bayes model were lower than those of both logistic regression and support vector machines. Nevertheless, the results were still commendable, with the model achieving a 94 percent accuracy across all evaluation metrics. However, it is notable that the number of misclassified samples using the Naive Bayes approach was significantly higher compared to the other two models. Despite this drawback, the model exhibited an impressive area under the curve, approaching one with a score of 0.99.

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	94.49	94.48	94.48	94.48

Table 6: Performance of Naive Bayes on fake news detection with TF-IDF word representation



(a) Confusion Matrix (Naïve Bayes)

(b) ROC Curve (Naïve Bayes)

Figure 8: Performance evaluation of Naïve Bayes using confusion matrix and ROC curve.

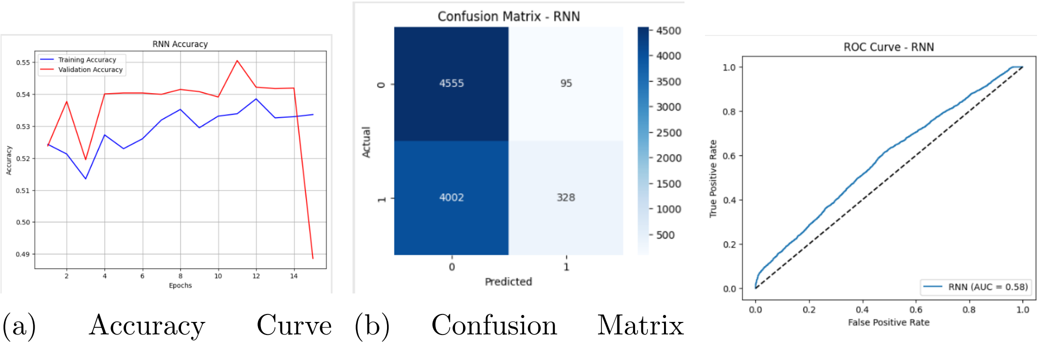
4.3 Sequential Models

4.3.1 RNN

In the context of the project, the Recurrent Neural Network emerged as the least effective model, demonstrating particularly low evaluation metrics, notably concerning the recall score. This model exhibited a pronounced bias towards predicting the ”Fake” class consistently, leading to a substantial number of false negatives. Additionally, the Receiver Operating Characteristic curve analysis revealed an Area Under the Curve value of 0.58, indicating that the model’s performance was nearly equivalent to random guessing. The aforementioned issues can be attributed to the inherent limitations of recurrent neural networks concerning their capacity to retain contextual information over extended sequences. As a result, these phenomena were anticipated based on the understanding of RNN dynamics.

Model	Accuracy	Precision	Recall	F1-Score
RNN	54.38	77.54	7.575	13.802

Table 7: Performance of RNN on fake news detection



(a) Accuracy Curve (RNN)

(b) Confusion Matrix (RNN)

(c) ROC Curve (RNN)

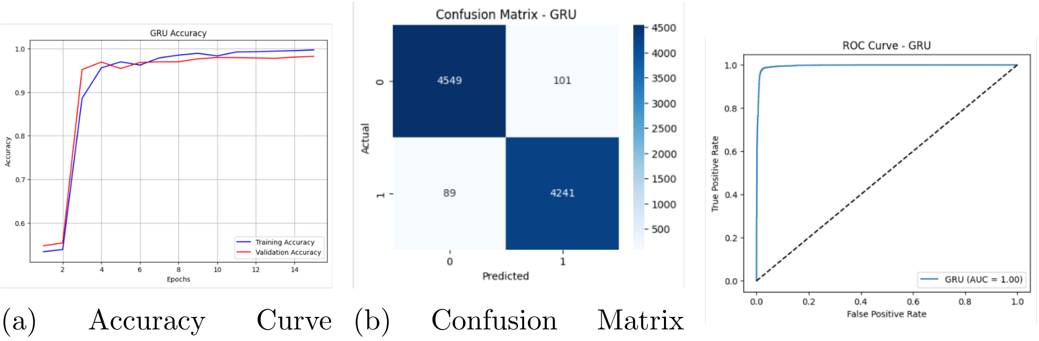
Figure 9: Performance evaluation of RNN model: accuracy curve, confusion matrix, and ROC curve.

4.3.2 GRU

This model demonstrated superior evaluation metrics compared to sequential models, surpassing the performance of LSTM, which was somewhat unexpected. Additionally, the model achieved an AUC score of 1, indicating perfect discrimination. It also exhibited nearly equal rates of false negatives and false positives, highlighting its balanced predictive capability.

Model	Accuracy	Precision	Recall	F1-Score
GRU	97.88	97.674	97.945	97.809

Table 8: Performance of GRU on fake news detection



(a) Accuracy Curve (GRU)

(b) Confusion Matrix (GRU)

(c) ROC Curve (GRU)

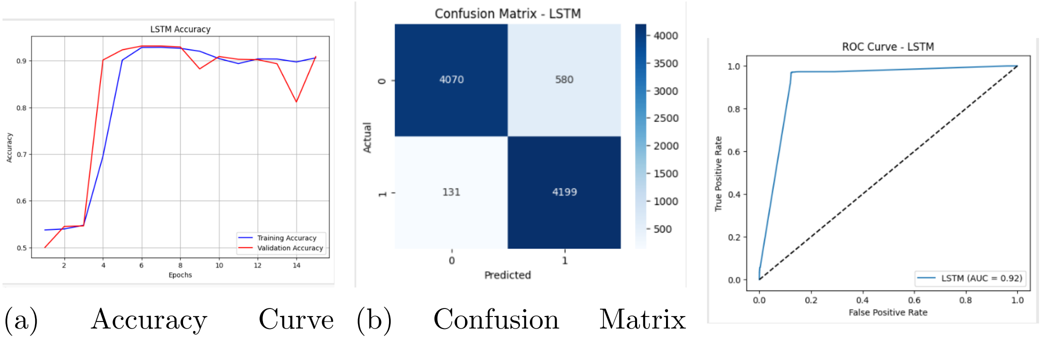
Figure 10: Performance evaluation of GRU model: accuracy curve, confusion matrix, and ROC curve.

4.3.3 LSTM

In the context of Long Short-Term Memory networks, the performance metrics revealed that while the overall scores remained commendably high, there was a notable decrease in the precision score. This decline indicates a propensity for the model to classify instances as 'Real' more frequently than warranted, leading to an increased rate of false positives. Nevertheless, the Area Under the Curve demonstrated strong performance, achieving a value of 0.92, which is approaching the optimal score of 1. This suggests that despite the challenges with precision, the model retains a robust ability to distinguish between classes effectively.

Model	Accuracy	Precision	Recall	F1-Score
LSTM	92.08	87.86	96.98	92.195

Table 9: Performance of LSTM on fake news detection



(a) Accuracy Curve (LSTM)

(b) Confusion Matrix (LSTM)

(c) ROC Curve (LSTM)

Figure 11: Performance evaluation of LSTM model: accuracy curve, confusion matrix, and ROC curve.

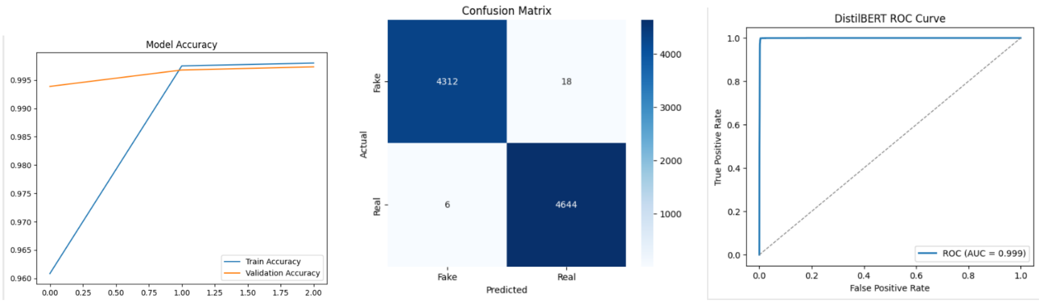
4.4 Transformers fine tuning

4.4.1 DistilBert

The model in question represents a distilled or lightweight variant of BERT, yet it demonstrates performance that surpasses all other models with the exception of BERT itself. The difference in performance between this distilled model and BERT is minimal, with both exhibiting exceptionally high scores. Notably, the occurrence of false negatives and false positives is remarkably low. This is evident in the perfect precision and recall scores of 100 percent. Furthermore, the area under the receiver operating characteristic curve was measured at 0.999, indicating an exceptionally high level of diagnostic accuracy.

Model	Accuracy	Precision	Recall	F1-Score
DistilBert	99.73	100	100	100

Table 10: Performance of DistilBert on fake news detection



(a) Accuracy (b) Confusion Matrix (c) ROC Curve

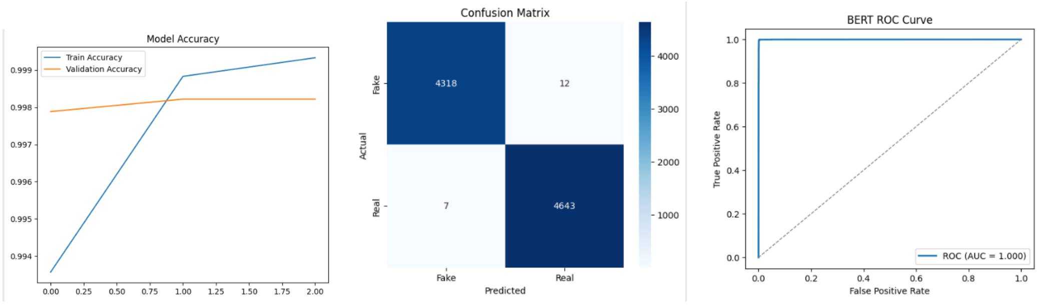
Figure 12: DistilBERT Results: Accuracy, Confusion Matrix, and ROC Curve.

4.4.2 Bert

The final model evaluated in our experiments is BERT (Bidirectional Encoder Representations from Transformers), which, as anticipated, outperformed all other models, including classical and sequential approaches. It achieved perfect scores across all metrics, with an area under the curve of 1.0. Additionally, the accuracy of the model was nearly perfect, registering at 99.79 percent. These results underscore the efficacy of transformer-based encoders in comprehending the subtleties and underlying meanings of textual data, thereby highlighting the superiority of BERT in natural language processing tasks.

Model	Accuracy	Precision	Recall	F1-Score
Bert	99.79	100	100	100

Table 11: Performance of Bert on fake news detection



(a) Accuracy (b) Confusion Matrix (c) ROC Curve

Figure 13: BERT Results: Accuracy, Confusion Matrix, and ROC Curve.

5 Ethical concerns

When discussing the detection of fake news, it is essential to consider a range of ethical implications that arise. Firstly, while the primary objective of such detection systems is to mitigate the spread of misinformation, there is a significant risk of bias in the classification processes. Certain demographic groups, topics, or linguistic contexts may be disproportionately flagged, leading to the potential silencing of legitimate voices and the reinforcement of existing imbalances in media representation. Additionally, the principle of transparency presents a critical concern. If these models operate as "black boxes," users may remain oblivious to the underlying rationale for a specific classification, which could erode trust in the system overall. Furthermore, the broader societal ramifications of determining what constitutes "fake" news must be carefully considered. The authority to define misinformation carries the potential for misuse, potentially serving to stifle dissenting opinions instead of safeguarding objective truth. Thus, it is imperative that the development and implementation of fake news detection systems are conducted with a strong emphasis on ethical considerations, fairness, and accountability.

6 conclusion

This project provided a comparative exploration of classical, sequential, and transformer-based approaches for fake news detection. The results show that while classical models, particularly when combined with TF-IDF, achieved near-perfect accuracy, their performance was largely dependent on surface-level lexical patterns rather than deeper semantic understanding. Sequential models demonstrated more variability, with GRU delivering strong results while RNNs performed poorly, highlighting the limitations of architectures that struggle to preserve long-range dependencies. By contrast, transformer-based models consistently outperformed all other approaches, underscoring their capacity to capture semantic nuances and contextual subtleties with exceptional precision. Beyond model performance, this study also emphasized the ethical dimensions of misinformation detection. Issues of bias, transparency, and definitional authority remain critical considerations, as technical effectiveness alone cannot guarantee fairness or trust. Collectively, the findings underscore both the potential of advanced NLP methods in addressing the challenges of fake news and the necessity of deploying such systems with caution and ethical awareness.