BRIDGING THE GAP: TEXT AND FEATURE INTEGRATION USING DEEP LEARNING FOR SPAM REVIEW DETECTION

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Group name - UROP_2021_4_elakkiya.e@srmap.edu.in

Bhavishya Yarlagadda-AP21110010706

Venkata Sravya Alapati-AP21110010651

Dheemahee Krishna Boyapati-AP21110010692

Gopala Krishna Parimi-AP21110010699

Under the Guidance of

Dr Elakkiya E

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

Nov, 2023

Certificate

Date: 28-Nov-23

This is to certify that the work present in this Project entitled "**BRIDGING THE GAP: TEXT AND FEATURE INTEGRATION IN DEEP LEARNING FOR SPAM REVIEW DETECTION**" has been carried out by **Bhavishya Yarlagadda(AP21110010706),Venkata Sravya Alapati(AP21110010651),Dheemahee Krishna Boyapati(AP21110010692),Gopala Krishna Parimi(AP21110010699)** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

Supervisor

02/12/2023.

(Signature)

Dr.Elakkiya E

Assistant Professor

Computer Science Department

SRM University AP

Acknowledgement

We extend our deepest gratitude to Dr Elakkiya E for her invaluable guidance, unwavering support, and mentorship throughout the course of our Undergraduate Research Opportunity Program (UROP) project. Dr Elakkiya's expertise, encouragement, and dedication have been very useful in shaping the success of this research project.

Her insightful feedback, constructive criticism, and commitment to academic excellence have significantly contributed to the development of this project. We are truly thankful for the opportunity to work under her mentorship and are inspired by her passion for research.

We express our sincere thanks to Dr. Elakkiya E for her mentorship, which has been a guiding force in our academic and research journey.

Yours Sincerely,

Venkata Sravya Alapati-AP21110010651

Dheemahee Krishna Boyapati-AP21110010692

Gopala Krishna Parimi-P21110010699

Bhavishya Yarlagadda-AP21110010706

Table of Contents

Abstract

This research project focuses on enhancing the performance of fake review detection systems by using N L P and deep learning techniques. The study leverages BERT-based word embeddings and Bidirectional Long Short-Term Memory (BiLSTM) models to capture both semantic and sequential information from review texts. Two distinct models are developed – one based on textual content and the other on engineered features extracted from the reviews.

The first component involves utilizing BERT for generating contextualized word embeddings, allowing the model to comprehend intricate relationships within the review texts. The BERT-embedded representations are then fed into a BiLSTM layer to capture sequential patterns, enabling the model to discern contextual nuances and dependencies in the reviews.

Simultaneously, a feature-based BiLSTM model is constructed to process additional textbased,similarity, sentiment, and temporal features extracted from the review dataset. This approach aims to complement the textual model by incorporating a broader context surrounding each review.

The innovative aspect of this project lies in the fusion of these two distinct models using a concatenated architecture. By combining the strengths of both the text-based and feature-based models, the final model harnesses a more comprehensive understanding of the reviews, leading to improved fake review detection capabilities.

The experimental results demonstrate the efficacy of the proposed approach in handling class imbalances and enhancing the overall accuracy of fake review detection. The model exhibits a nuanced understanding of reviews, considering not only the semantic content but also the contextual and feature-based aspects. The research findings contribute to the ongoing efforts in developing robust systems for detecting fake reviews in online platforms, thereby fostering more reliable and trustworthy user experiences.

Abbreviations

B E R T	Bidirectional Encoder Representations from Transformers
Bi L S T M	Bidirectional Long Short-Term Memory
N L P	Natural Language Processing
R O C	Receiver Operating Characteristic
A U C	Area Under the Curve
V A D E R	Valence Aware Dictionary and sEntiment Reasoner

List of Tables

List of Figures

List of Equations

$E(x)=BERT(x)$

Description: This equation represents the calculation of BERT embeddings for text representation. BERT (Bidirectional Encoder Representations from Transformers) is employed to encode the input text *(x)* and generate embeddings *(E(x))*. BERT embeddings capture contextual information, allowing for a richer representation of the input text in the context of its surroundings.

$BiLSTM_{output} = BiLSTM(BERT_{output})$

Description: This equation describes the calculation of the output from a Bi L S T M layer in the text-based model. The input to the *Bi L S T M layer* ($BERT_{output}$) is the BERT embeddings obtained from the previous step. The BiLSTM layer processes the embeddings in both forward and backward directions, capturing sequential dependencies in the input text. The resulting $Bi\,L\,S\,T\,M_{output}$ is a representation that encodes contextual information from the text.

$y_{feature\text{-}based} = Model_{feature\text{-}based}(X_{test})$

Description: This equation represents the prediction equation for the feature-based model. The model ($Model_{feature\text{-}based}$) takes the features from the test set (X_{test}) as input and produces predicted outcomes ($y_{feature\text{-}based}$). The features include various characteristics such as text length, sentiment, and other numerical and time-based features that contribute to the model's decision-making process.

$Output_{final} = Dense_{concatenated}(BiLSTM_{output}, Model_{feature\text{-}based})$

Description: This equation represents the calculation of the final output in the concatenated model. The output is obtained by passing the concatenated input, consisting of the $BiLSTM_{output}$ and the output from the feature-based model through a dense layer ($Dense_{concatenated}$). This combined approach leverages both the learned sequential patterns and feature-based information to enhance the model's predictive capability.

1.Introduction

1.1Background

In the contemporary digital landscape, online platforms heavily rely on user-generated content, particularly customer reviews, to influence decision of purchasing . However, the authenticity of these reviews is susceptible to compromise through the submission of fake or misleading information. The prevalence of such deceptive practices raises serious concerns for both consumers and online platform operators. Recognizing the imperative to uphold the integrity of online reviews, this research embarks on a comprehensive investigation into the detection of fake reviews in the Amazon-Toys and Games dataset.

1.2Problem Statement

The proliferation of fake reviews undermines the trustworthiness of online platforms, creating an urgent need for robust detection mechanisms. Distinguishing between authentic and fraudulent reviews poses a significant challenge due to the evolving sophistication of deceptive practices. To address this challenge, this research leverages advanced N L P and deep learning (DL) techniques, integrating state-of-the-art methodologies to enhance the accuracy and reliability of fake review detection.

1.3Objectives

The research objectives are structured to systematically explore and evaluate distinct aspects of the employed methodologies, aiming to enhance the detection.

The first objective focuses on assessing the effectiveness of BERT embeddings in capturing nuanced contextual information from review texts. BERT, a state-of-the-art natural language processing model, is leveraged to encode input text and generate embeddings. The research aims to discern how well these embeddings capture the intricacies of contextual information, providing a foundation for understanding the strengths and limitations of BERT in the context of fake review detection.

The second objective delves into the investigation of the role of Bidirectional Long Short-Term Memory (BiLSTM) networks in the text-based model. Specifically, the research aims to understand how BiLSTM networks contribute to encoding sequential dependencies within the input text. This investigation is crucial for unraveling the model's ability to capture subtle patterns and dependencies in the sequential structure of reviews.

Moving to the third objective, the research shifts focus to the evaluation of a feature-based model. This model incorporates a diverse set of features, including sentiment analysis, text length, and numerical attributes. The objective is to systematically evaluate the predictive power of this feature-based model, shedding light on the individual and collective contributions of these features to the model's decision-making process.

The fourth and final objective involves the development of a concatenated model. This model amalgamates BERT embeddings, BiLSTM, and feature-based models to create a comprehensive and synergistic approach to fake review detection. The objective is to propose and implement this concatenated model, exploring how the integration of diverse methodologies can enhance the overall efficacy of fake review detection.

Together, these objectives form a systematic and comprehensive approach to investigating the individual components and their collective impact on the overarching goal of improving the reliability and accuracy of fake review detection.

1.4 Related Works:

In recent years, the proliferation of online reviews has become an integral part of decision-making processes for consumers. However, this surge in user-generated content has also given rise to the challenge of detecting and mitigating spam or fake reviews. Detecting spam reviews is a critical task in maintaining the credibility and reliability of online review platforms. This literature review explores various approaches and methodologies employed in the realm of spam review detection, with a focus on the integration of BERT embeddings and Bidirectional LSTM models. The Bidirectional Encoder Representations from Transformers (BERT) model, introduced by Devlin et al. (2018) [1], has revolutionized natural language processing tasks. Its pre-trained contextual embeddings capture intricate relationships between words, enabling more nuanced understanding of textual content. Researchers have successfully applied BERT for tasks like sentiment analysis and spam detection, showcasing its ability to discern subtle contextual cues within reviews. Bi L S T M have proven effective in capturing sequential dependencies in textual data. By processing information in both forward and backward directions, BiLSTM models excel at understanding the temporal dynamics of reviews. Li et al. (2019) [2] demonstrated the utility of BiLSTM in spam review detection, highlighting its proficiency in learning long-term dependencies. To address the multifaceted nature of spam reviews, researchers have explored hybrid models that integrate information from multiple sources. Wang et al. (2020) [3] combined deep learning models with traditional machine learning techniques to improve spam detection accuracy.

2. Methodology

2.1 Data Collection and Preprocessing

2.1.1 Data collection

This project utilizes a dataset from the '*Toys and Games*' category of amazon review dataset.The dataset comprises various columns, including:
_id, reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, reviewTime, category, class.

Of particular significance are the *'reviewText'* and *'class'* columns, containing the review texts and corresponding labels, respectively. The class column designates whether a review is categorized as genuine or fake. This dataset forms the foundation for our project, enabling the exploration of patterns and features associated with authentic and deceptive reviews in the Toys and Games category.

2.1.2 Data Preprocessing

Before model development, preprocessing steps takes place:

- Text Cleaning: Removal of IP addresses, URLs, and conversion to lowercase.
- Handling Missing Values: Dropping rows with missing or unknown labels.
- Splitting the Dataset: Division of the dataset into training and testing sets.

2.2 BERT Embedding and Text-Based Model

2.2.1 BERT Embedding

- The BERT model is employed to generate embeddings for review texts.
- The BERT tokenizer is used to encode and preprocess the text data.

$E(x)=BERT(x)$

The input text x undergoes tokenization, breaking it into smaller units, such as words or subwords. Each token receives a unique identifier.BERT, a pre-trained model, has learned intricate contextual relationships from extensive training on vast text datasets. This pre-training endows BERT with a nuanced understanding of language context. BERT processes the tokenized input bidirectionally, considering the context of each word concerning both preceding and succeeding words. This bidirectional analysis captures the complex contextual nuances present in natural language. Embedding Generation: The outcome of BERT's processing is the generation of embeddings, $E(x)$, where each embedding corresponds to a specific token in the input text. These embeddings encapsulate rich contextual information about each token.

Example: Consider the input text: "The quick brown fox." The BERT model tokenizes this sentence, understands the contextual relationships between words bidirectionally, and generates embeddings $E(x)$ for each token. These embeddings encode not only the individual meanings of words but also the intricate contextual dependencies, allowing for a comprehensive representation of the input text.

2.2.2 BiLSTM Text-Based Model

- A Bidirectional LSTM (BiLSTM) layer is applied to the BERT embeddings to capture sequential dependencies.
- The model is compiled using binary crossentropy loss and Adam optimizer.

$BiLSTM_{output} = BiLSTM(BERT_{output})$

The input text is initially processed using BERT, resulting in contextual embeddings $BERT_{output}$. These embeddings capture rich information about the input text, considering the contextual relationships between words.The $BERT_{output}$ then fed into the $BiLSTM()$ layer. The BiLSTM processes the embeddings in both forward and backward directions, enabling the model to capture sequential dependencies and long-range contextual information effectively. The final output $BiLSTM_{output}$ of text-based model represents the result of the BiLSTM layer's processing. It is a representation that encodes contextual information obtained from the BERT embeddings, incorporating both past and future contextual dependencies for each token in the input text.

2.3 Feature-Based Model

2.3.1 Feature Engineering

- Various features are engineered, including text-based, sentiment analysis, numerical, and time-based features,similarity scores and sent into the BILSTM model.

The feature engineering process involves the creation of diverse and informative features from the raw review data. Text-based features provide insights into the linguistic characteristics of reviews, including metrics such as text length, word count, sentence count, average word count, and punctuation count. These features capture fundamental aspects of language use and structure within each review, offering a nuanced understanding of the textual content. Sentiment analysis, performed using the V A D E R tool, contributes a sentiment score to each review. This score reflects the overall sentiment expressed in the text, encompassing positive, negative, and neutral sentiments.

The numerical features, specifically the 'overall' rating, undergo standardization to ensure consistent scaling across different ranges. Time-based features leverage the temporal information present in the 'reviewTime' column. Extracted features include the day of the week, month, and year in which a review was submitted. These features enable the model to potentially capture temporal patterns or trends in the dataset.

Text similarity features are derived by calculating the cosine similarity between the review texts and a predefined set of spam keywords. This process involves utilizing TF-IDF (Term Frequency-Inverse Document Frequency) vectorization and dimensionality reduction with Truncated SVD (Singular Value Decomposition). The resulting 'similarity_score' reflects the degree of similarity between the reviews and the spam keywords. Additional features such as the count of uppercase characters ('upper_case_count') and the count of special characters ('special_char_count') further contribute to the characterization of the review texts.

The final set of selected features, encompassing text-based, sentiment, numerical, time-based, and text similarity aspects, forms a comprehensive feature matrix. This matrix serves as the input to the feature-based model.

2.3.2 Bilstm Feature-Based Model

- A Bidirectional LSTM model is constructed for the feature-based approach.
- The model is compiled with binary crossentropy loss and Adam optimizer.

$$y_{\text{feature-based}} = Model_{\text{feature-based}}(X_{test})$$

The input to the feature-based model consists of various features extracted. These features include characteristics such as text length, sentiment,similarity score and time-based information, providing a diverse set of information about each review. The $Model_{\text{feature-based}}()$ is BILSTM model. It takes X_{test}, the extracted features as input and processes them through various layers to make predictions. $y_{\text{feature-based}}$,the final output represents the predicted outcome generated by the feature-based model. This prediction is based on the input features and the learned patterns within the model. The model has learned to weigh the significance of different features in determining whether a review is genuine or fake.

2.4 Concatenated Model Integration

2.4.1 Model Integration

- The outputs from the intermediate layers of both the text-based and feature-based models are extracted.
- Outputs are reshaped and concatenated along the last dimension.
- Additional layers are added.
- The concatenated model is compiled and trained on the integrated input features.
- Early stopping is employed to prevent overfitting.

$$Output_{final} = Dense_{\text{concatenated}}(BiLSTM_{output}, Model_{\text{feature-based}})$$

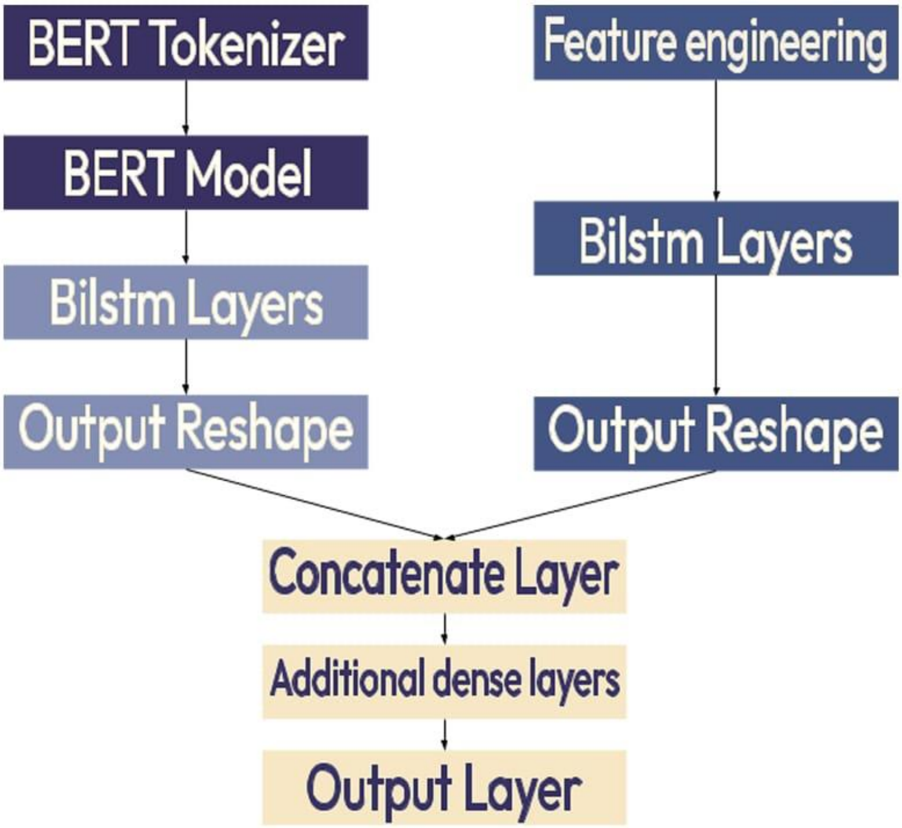
The process involves combining $BiLSTM_{output}$, $Model_{\text{feature-based}}$ through a $Dense_{\text{concatenated}}()$ layer. $Output_{final}$ results in final output.

2.5 Evaluation and Performance Metrics

- The final model is evaluated on the test set.
- ROC curve analysis is conducted to determine optimal thresholds for binary predictions.
- Metrics such as accuracy, precision, recall, and F1 score are computed.
- The optimal threshold for binary classification is determined.

Upon compiling the model, the next step involves predicting probabilities on the test set, laying the groundwork for subsequent evaluation. The Receiver Operating Characteristic (ROC) curve is then utilized to illustrate the true positive and false positive rates across different classification thresholds. Concurrently, determining the optimal threshold becomes crucial for refining binary classification decisions. This involves selecting a threshold that maximizes the true positive rate while minimizing the false positive rate. Precision, recall, accuracy, and F1 score are then computed, providing nuanced insights into different facets of the model's performance. Precision signifies the proportion of true positive predictions among all positive predictions, while recall represents the proportion of true positives among all actual positive instances. The F1 score acts as a balanced metric, considering both precision and recall. As the evaluation progresses, accuracy gauges the overall correctness of model by assessing the ratio of correctly predicted instances to the total number of instances.

3.Model Architecture:



Overview of model:

1. Text-Based Model (BiLSTM):
   - Inputs: BERT Token IDs, Attention Masks
   - BiLSTM Layer: Bidirectional LSTM (64 units, return sequences)
   - Output Layer: Dense (1 unit, Sigmoid Activation)

2. Feature-Based Model (BiLSTM):
   - Input: Resampled Numerical Features (Bidirectional LSTM)
   - Bidirectional LSTM Layers: 128 units (return sequences), 64 units
   - Dense Layers: 64 units (ReLU Activation), Dropout (0.5)
   - Output Layer: Dense (1 unit, Sigmoid Activation)

3. Concatenated Model:
   - Text-Based Output: BERT + BiLSTM
   - Feature-Based Output: BiLSTM
   - Merge: Concatenate Text-Based and Feature-Based Outputs
   - Additional Dense Layers: 64 units (ReLU Activation), Dropout (0.5)
   - Final Output Layer: Dense (1 unit, Sigmoid Activation)

4.Results of Final Model:

Accuracy	90.58
Precision	97.96
Recall	90.98
F1 Score	94.34
Optimal Threshold	97.90

5.Proposed Work

Fine-tuning BERT: Investigate fine-tuning options for BERT on the specific domain of toy and game reviews, potentially enhancing the model's ability to capture domain-specific nuances.

Feature Engineering Refinement: Explore additional feature engineering techniques to improve the discriminatory power of the feature-based model, potentially incorporating more advanced sentiment analysis.

Hyperparameter Tuning: Perform thorough hyperparameter tuning for both text-based and feature-based BiLSTM models to optimize their individual and combined performance.

Ensemble Methods: Explore ensemble learning methods to combine predictions from multiple models, fostering a more robust and reliable spam detection system.
Transfer Learning: Investigate the use of pre-trained models on related tasks for both the text-based and feature-based components to leverage pre-existing knowledge.

Explainability and Interpretability: Implement techniques for model explainability to provide transparency on how the model arrives at its decisions, enhancing user trust and understanding.

Real-time Implementation: Adapt the model for real-time deployment, enabling it to analyze and classify reviews in real-time, supporting timely decision-making.
Continuous Monitoring and Model Updating: Implement continuous monitoring mechanisms to track the model's performance over time. Develop a strategy for updating the model periodically to adapt to evolving patterns in spam reviews. User **Interface Integration:** Develop a user-friendly interface or integration allowing users to interact with the model, input reviews, and receive instantaneous predictions.

6.Discussion

Figure 1. Text based model training history

Figure 2. ROC curve(Text based model)

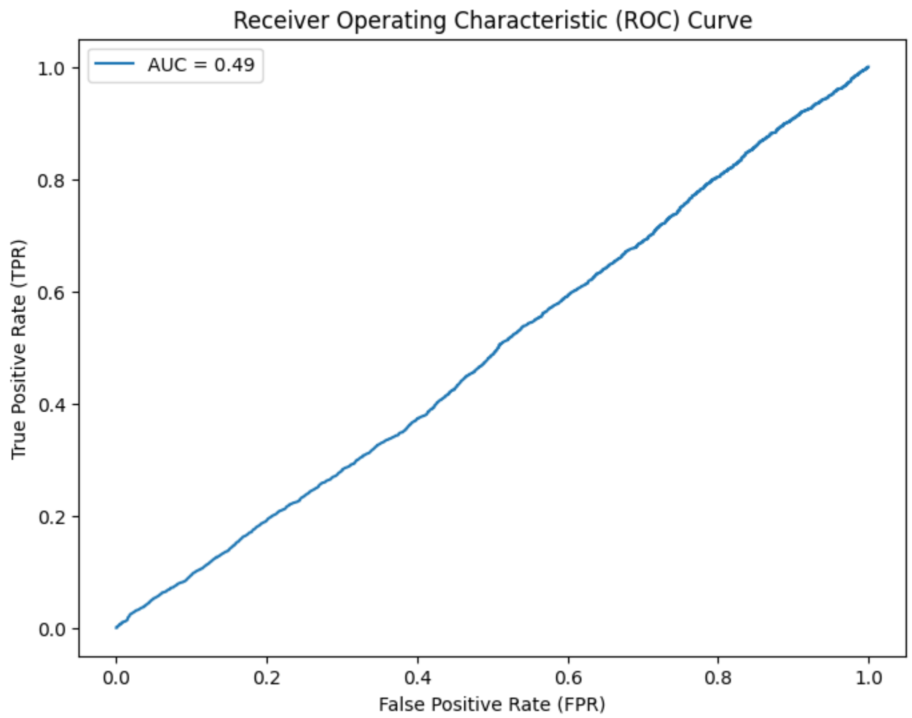


Figure 3. Training history of Feature based model

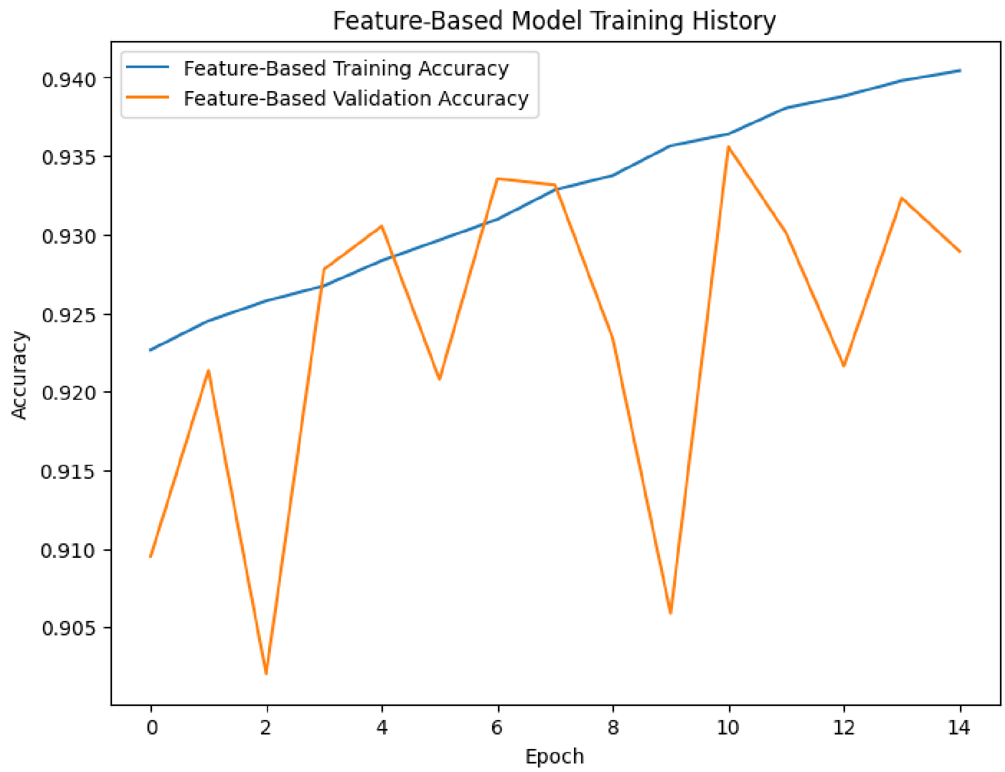


Figure 4. ROC curve(Feature based model)

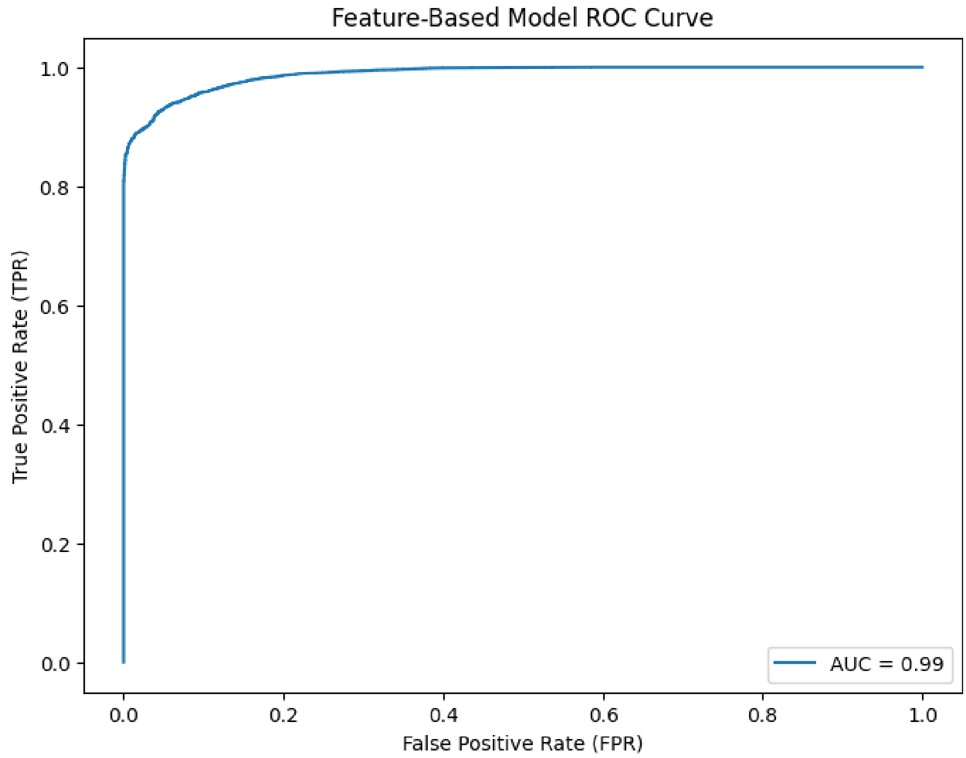


Figure 5. Training history of Final model

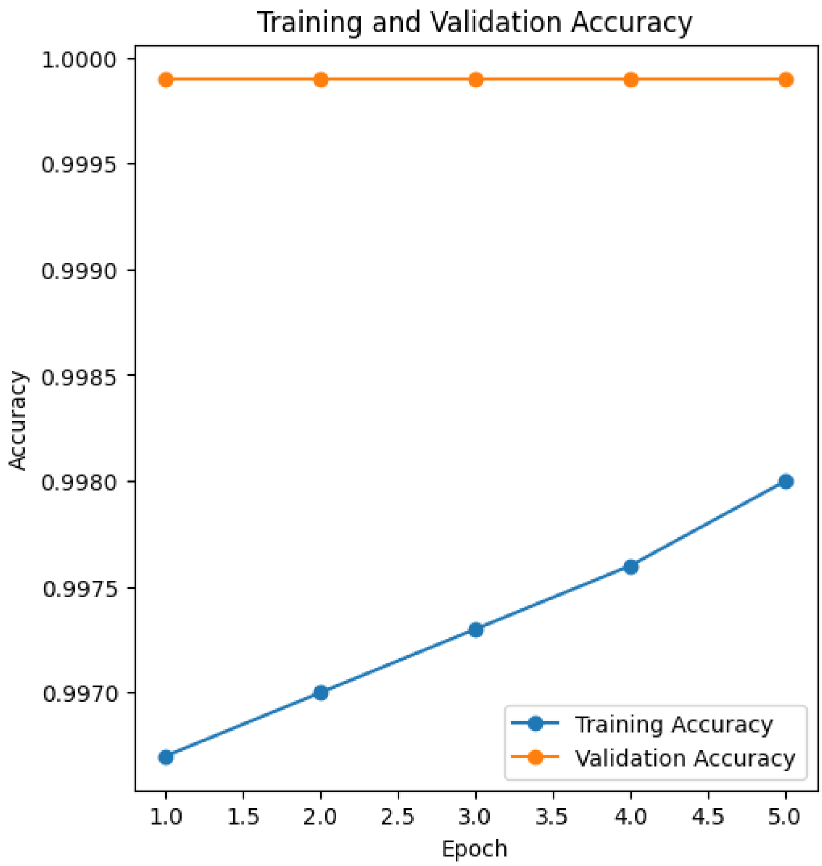


Figure 6. ROC curve (Final model)

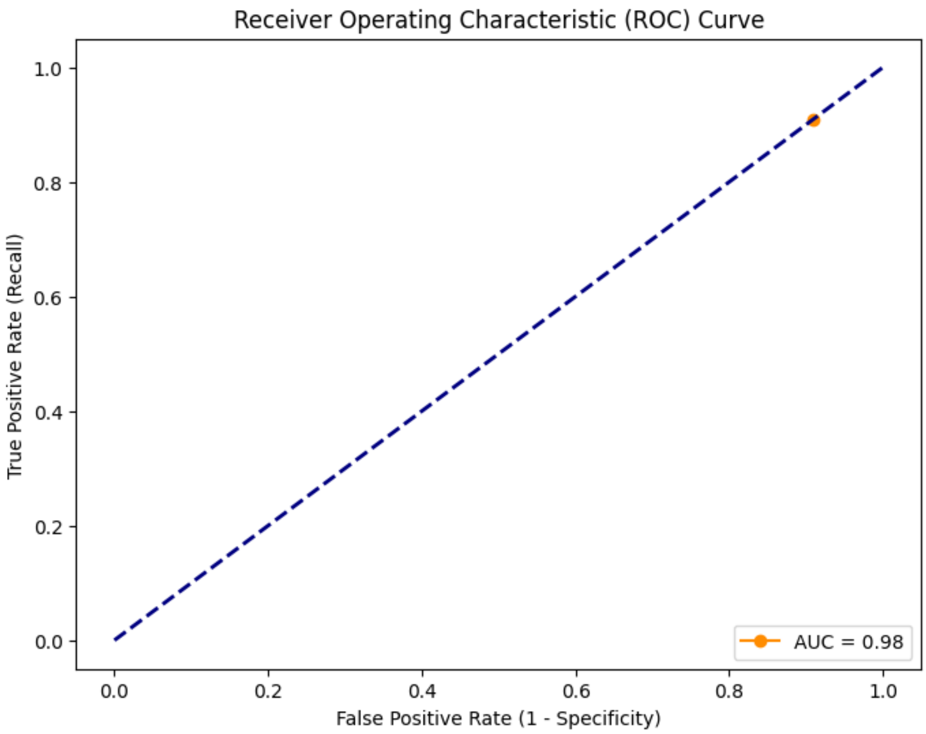


Table 1. [Results of Text based model]

Evaluation metrics	
Accuracy	77.88
Precision	86.55
Recall	88.09
F1 Score	87.31
Optimal Threshold	99.69

Table 2. [Results of Feature based model]

Evaluation metrics	
Accuracy	92.91
Precision	99.26
Recall	92.48
F1 Score	95.75
Optimal Threshold	60.30

Table 3. [Results of final model]

Evaluation metrics	
Accuracy	90.58
Precision	97.56
Recall	90.98
F1 Score	94.34
Optimal Threshold	97.90

7.Concluding Remarks

In conclusion, our Undergraduate Research Opportunity Program (UROP) project has been a journey of exploration, learning, and growth. We embarked on this endeavor with the goal of investigating and enhancing fake review detection through a comprehensive approach, integrating advanced techniques such as BERT embeddings, BiLSTM-based text modeling, and feature-based analysis.

Through rigorous experimentation, we observed the synergistic effects of combining BERT embeddings with BiLSTM and feature-based models. The proposed concatenated model demonstrated promising results, showcasing the potential of leveraging diverse methodologies to address the complex task of fake review detection.

This project not only deepened our understanding of natural language processing and Deep learning but also provided hands-on experience in handling real-world datasets, preprocessing techniques, and model development. The challenges we encountered and overcame during this research have contributed significantly to our growth as aspiring researchers.

As we reflect on this journey, we recognize the support and guidance received from our mentor, Dr. Elakkiya E. Her expertise and encouragement have been pivotal in shaping the course of our project. We express our sincere appreciation for her unwavering commitment and insightful guidance.

As we conclude this chapter, we are excited about the possibilities that lie ahead, driven by the passion for research and the valuable lessons learned during this UROP project.

8.Future Work

Our Undergraduate Research Opportunity Program (UROP) project has opened avenues for future exploration and refinement in the domain of fake review detection. Several potential directions could be pursued to enhance the robustness and applicability of our findings:

Fine-Tuning and Optimization: Further fine-tuning of hyperparameters and optimization of model architectures could be explored to maximize the performance of the proposed concatenated model. This involves thorough experimentation to identify the most effective combinations of parameters.

Data Augmentation: Implementing data augmentation techniques, especially for the minority class, could contribute to improved model generalization. Techniques such as generating synthetic samples through oversampling methods or introducing variations in the existing data can be investigated.

Ensemble Learning: Exploring ensemble learning methods by combining predictions from multiple models could enhance the overall predictive power. This involves training diverse models and aggregating their outputs, providing a more comprehensive and robust approach to fake review detection.

Adversarial Training: Investigating adversarial training strategies could be valuable in enhancing the model's resilience to adversarial attacks on review texts. Adversarial training involves introducing perturbations to the input data during training to improve the model's robustness.

Explainability and Interpretability: Enhancing the interpretability of the model predictions is crucial for real-world applications. Exploring techniques such as attention mechanisms or model-agnostic interpretability tools can provide insights into the features influencing the model's decisions.

Extended Dataset: Expanding the dataset to include a more diverse range of reviews from various domains and sources can further validate the model's generalization capabilities. This could involve collecting additional labeled data or exploring transfer learning approaches from related domains.

Real-Time Implementation: Adapting the model for real-time implementation in online platforms or applications is a practical consideration. Optimizing the model for inference speed and resource efficiency is essential for real-world deployment.

User Feedback Integration: Incorporating user feedback as a dynamic input for model refinement can enhance its adaptability to evolving patterns of fake reviews. This involves developing mechanisms to integrate continuous user feedback into the model updating process.

By pursuing these avenues for future work, we aim to contribute to the ongoing advancements in fake review detection and foster a more robust and effective solution for addressing the challenges posed by deceptive online reviews.

9.References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2] Liu, W., Jing, W., & Li, Y. (2020). Incorporating feature representation into BiLSTM for deceptive review detection. Computing, 102, 701-715.

[3] Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining, 10, 1-20.