UNVEILING ONLINE DECEPTION: A DEEP LEARNING APPROACH TO DETECT FAKE REVIEWS

Masters in Data Science

Research Thesis Report

Liverpool John Moore University

Sridhar Pai Tonse

STUDENT ID: 1096325

THESIS SUPERVISOR : DR. ANIL VUPPALA

March 2024

Dedication

To Ancient Saints and Divine Power, though which all our knowledge originated and laid the foundation for all the new knowledge systems and for the opportunities I hold today. May this work expand their legacy by contributing to the betterment of human lives. And to the indomitable human spirit. a constant source of inspiration in our pursuit of knowledge and understanding. May this research contribute to the journey of continuous evolution of human being for better future.

Acknowledgements

I want to express my sincere gratitude and appreciation to my Thesis Supervisor Dr. Anil Vuppala for his valuable support during my Master’s Thesis. Despite my many reasons for lack of productivity, he always kept his optimism high, guided me with ideas, and helped move me forward. Most importantly, his guidance helped maximize my project’s key elements. I am grateful for the time I have spent with him as it has helped my interest in particular fields of Deep learning and NLP.

Abstract

In the era of digital commerce, online reviews significantly influence consumer behaviour and business reputation. But we can’t always trust these reviews because some of them might be fake. This thesis delves into the use of sophisticated Deep Learning methods to accurately identify fake reviews. We propose a novel framework that harnesses the power of Natural Language Processing (NLP) and Deep Learning to analyse the linguistic patterns and sentiment of online reviews. Our approach includes the use of feature extraction methods and classification algorithms to distinguish between genuine and deceptive reviews. The proposed method will leverage machine learning algorithms and deep learning techniques such as Logistic Regression, SVM, Random Forest, along with deep learning networks such as CNN, RNN-LSTM and BERT. In addition, delves into traditional and deep learning techniques of embeddings. This thesis findings confirms the deep learnings techniques such as BERT embeddings along with deep learning model can provide much superior performance compared to traditional tokenizers such as TFIDF. The results demonstrate the effectiveness of our proposed model in accurately identifying fake reviews, thereby contributing to the integrity of online platforms and assisting users in making informed decisions. This research opens up new avenues for enhancing the reliability of online review systems and has significant implications for businesses, consumers, and policy-makers in the digital marketplace. The research will also draw insights from a comprehensive review of around 40 research papers, comparing various techniques used for detection of spam reviews.

Table of Contents

Dedication ii

Acknowledgements iii

Abstract iv

List of Figures vii

List of Tables vii

CHAPTER 1: INTRODUCTION 1

1.1 Background of the Study 1

1.1.1 Content-Based Methods: 1

1.1.2 Behavior-Based Methods 1

1.1.3 Graph-Based Methods 1

1.1.4 Supervised Methods 2

1.1.5 Unsupervised Methods 2

1.1.6 Semi-Supervised Methods 2

1.1.7 Research Gap 2

1.1.8 Purpose of this Study 2

1.2 Problem Statement 3

1.3 Aim and Objectives 3

1.4 Research Questions 3

1.5 Scope of the Study 4

1.6 Significance of the Study 4

1.7 Structure of the Study 5

CHAPTER 2: LITERATURE REVIEW 7

2.1 Introduction 7

A Journey Through Fake Review Detection. 7

2.2 Early Approaches (Pre-2010s): 7

2.3 Machine Learning Approaches 8

2.4 Deep Learning Approaches 9

2.5 Summary 13

CHAPTER 3: RESEARCH METHODOLOGY 15

3.1 Introduction 15

3.2 Research Methodology Framework 15

3.2.1 Fake Review Detection Methodology Framework 15

3.2.2 Proposed Methodology 15

3.2.3 Data Sets: 16

3.2.4 Data Pre-Processing: 17

3.2.4 Embedding Techniques: 17
3.2.5 Class balancing 19
3.2.6 Model Training: 20
3.2.7 Model Evaluation: 20
3.2.8 Results Analysis: 21
3.3 Maching Learning Techniques 22
3.3.1 Logistic Regression 22
3.3.2 SVM (Support Vector Machine) 22
3.3.3 Naïve Bayes 23
3.3.5 Random Forest 24
3.3.6 XGBoost 26
3.4 Deep Learning Models. 27
3.4.1 Convolution Neural Networks (CNN) 27
3.4.2 Recurrent Neural Networks (RNN) 28
3.4.3 Long Short-Term Memory (LSTM) 28
3.4.4 RNN-LSTM 30
3.4.5 BERT (Bidirectional Encoder Representations from Transformers) 30
3.5 Summary 30
CHAPTER 4: IMPLEMENTATION 33
4.1 Introduction 33
4.2 Exploratory Data Analysis (EDA) 33
4.2.1 Data Set Description 33
4.2.2 Data definition of Yelp Labeled Data Set. 34
4.2.3 Distribution of target variable ie LABEL (Fake vs. Genuine) 34
4.2.6 Check the dimensions of the dataset 35
4.2.7 Sample Fake reviews vs Genuine Reviews 36
4.2.8 Word Cloud Fake Reviews vs Genuine Reviws 37
4.3 Resource Requirements 37
CHAPTER 5: RESULTS AND EVALUATION 39
5.1 Introduction 39
5.2.1 TFIDF Embedding Comparison with Machine Learning Classifiers 39
5.2.2 Glove Embedding Comparison with Machine Learning Classifiers 44
5.2.3 BERT based embedding Comparison with ML and DL Classifiers 50
5.2.4 Comparison of Deep Learning Classifiers 52
5.3 Model Prediction with Live Data: 55
5.3.1 Testing of Deep Learning Models on Live Yelp data set : 55
5.4 Summary 56

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS 57
REFERENCES 58
APPENDIX A: RESEARCH PLAN 60
APPENDIX B: RESEARCH PROPOSAL 62
APPENDIX C: ETHICS FORMS 79

List of Figures

Figure 1 Fake Review Detection Methodology Framework 15
Figure 3 A BERT based Sentence Transformers architecture (Reimers and Gurevych, 2019) 16
Figure 4 BERT Illustration 18
Figure 5 Support Vector Machine 22
Figure 6 Naive Bayes 23
Figure 7 Random Forest Illustration 25
Figure 8 CNN architecture 27
Figure 9 RNN Architecture 28
Figure 10 LSTM Architecture 29
Figure 11 Distribution of target variable ie Label 34
Figure 12 Distribution of Genuine Review Text Lenghts 35
Figure 13 Distribution of Fake Review Text Lenghts 35

List of Tables

Table 1Yelp Labelled Data Set Description 34
Table 2 TFIDF Embedding Technique comparison with ML Classifiers 44
Table 3 Glove Embedding Technique comparison with ML Classifiers 50
Table 4 BERT, RoBERTa, DistilBERT and XLNet Embedding comparison across ML/DL Models 52
Table 5 Deep Learning Classifiers Comparison 55
Table 6 Research Plan for the Thesis Fake Review Detection 60

List of Abbreviations

LR	Logistic Regression
RF	Randon Forest
NB	Naïve Bayes
SVM	Support Vector Machine
RNN	Recurrent Neural Network
CNN	Convolution Neural Network
LSTM	Long Short Term Memory
BERT	Bi Directional Encoder Representation Transformers
ROBERTA	A Robustly optimized BERT pre training Approach
DISTILBERT	A Distilled version of BERT
XLNET	
ROC	Receiver Operating Characteristic Curve (ROC Curve)
AUC	Area Under the ROC Curve

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Fake reviews are a serious problem that affects the trustworthiness and credibility of online platforms, such as e-commerce, social media, and travel websites. Fake reviews can mislead consumers, damage the reputation of businesses, and distort the market competition. Therefore, detecting and combating fake reviews is an important and challenging task that requires advanced machine learning techniques.

Existing methods for fake review detection can be broadly categorized into three types: content-based, behavior-based, and graph-based.

1.1.1 Content-Based Methods: These methods focus on the textual content of the reviews. They often involve extracting features from the text, such as n-grams, part-of-speech tags, and psycholinguistic features(Tang and Cao, 2020). Techniques like machine learning and deep learning are used to classify reviews as genuine or fake based on these features(Hajek et al., 2020). For instance, methods like Support Vector Machines (SVMs) and neural networks have been used for fake review detection(Hajek et al., 2020). Deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have also been proposed for this task(Singh et al., 2023).

1.1.2 Behavior-Based Methods: These methods consider the behavior of the reviewers. For instance, they may look at the timing of the reviews, the rating patterns, and other behavioral aspects (Wang Junrenand Chen, 2023). The idea is that fake reviewers may exhibit certain patterns in their behavior that can be used to identify them. For example, a model might look at the average cosine similarity of a reviewer's reviews, or it might use a Convolutional Neural Network (CNN) to extract patterns from the behavioral information (Wang Junrenand Chen, 2023).

1.1.3 Graph-Based Methods: These methods construct a graph where the nodes represent entities like reviewers, reviews, and products, and the edges represent relationships between these entities. The graph is then analyzed to detect fake reviews.(Fang et al., 2020) For instance, one approach might start with a small subgraph of known fake reviewers and then expand the subgraph by adding connected suspicious reviewers (Manaskasemsak et al., 2023). Another approach might use a dynamic knowledge graph, where time series related features are added to the graph construction process (Fang et al., 2020).

Also learning methods used for fake review detection can be classified into Supervised methods, Unsupervised methods and Semi-supervised methods as explained below.

1.1.4 Supervised Methods: These methods rely on labeled data, i.e., reviews that are already marked as genuine or fake. Machine learning algorithms are trained on this data to create a model that can classify new reviews. For instance, Support Vector Machines (SVMs), Naive Bayes, K-Nearest Neighbors (KNN), and logistic regression have been used for fake review detection(Elmogy et al., n.d.). Deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have also been proposed for this task(Mir et al., n.d.).

1.1.5 Unsupervised Methods: These methods do not require labeled data. Instead, they try to learn the underlying structure of the data or identify unusual patterns. For example, one approach might involve using a user's behavior representation rather than review contents to measure authenticity (Li Qianand Wu, 2019). Another approach might use Long Short-Term Memory (LSTM) networks and autoencoders to distinguish spam reviews from other real reviews.(Saumya and Singh, 2022)

1.1.6 Semi-Supervised Methods: These methods use a combination of labeled and unlabeled data. The idea is to use a small amount of labeled data to guide the learning process on a larger amount of unlabeled data. Techniques such as Co-training, Expectation Maximization, Label Propagation and Spreading, and Positive Unlabeled Learning have been proposed for fake review detection.(Swathi et al., 2023).(Rakibul and Md. Rabiul, 2019)

1.1.7 Research Gap:

- Existing studies in fake review detection predominantly center around supervised learning methods, which heavily rely on **Labeled data** for training and testing.
- There remains a significant gap in the exploration of **Embedding techniques** coupled with machine learning (ML) and deep learning (DL) approaches for fake review detection.
- Despite the growing interest in advanced Embedding Techniques such as **TF-IDF, Glove, BERT, ALBERTA, DISTILBERT, and XLNET**, their application in this domain remains relatively unexplored.
- This research aims to address this gap by investigating the effectiveness of these **Embedding techniques** in conjunction with ML and DL methods for fake review detection

1.1.8 Purpose of this Study is to delve into the utilization of various Embedding Techniques, including TF-IDF, Glove, BERT, ALBERTA, DISTILBERT, and XLNET, in tandem with machine learning (ML) and deep learning (DL) approaches for fake review detection. By focusing on these advanced tokenization methods, the study seeks to enhance the accuracy and robustness of fake review detection systems. Furthermore, this research endeavors to explore the potential of semi-supervised and unsupervised learning paradigms in conjunction with Embedding Techniques to mitigate the reliance on labeled data, thereby facilitating scalability and generalizability across diverse domains. Through empirical evaluation and comparative analysis, this study aims to contribute novel insights and methodologies to the field of fake review detection, ultimately advancing the state-of-the-art in detecting fraudulent reviews.

1.2 Problem Statement

1. The proliferation of fake reviews online presents a complex challenge for consumers, businesses, and digital platforms. Identifying of fake reviews is complex challenge.
2. This research aims to come up with novel approach to identify fake reviews. As part of this we will study and compare various Embedding Techniques for fake review detection such as **TF-IDF, Glove, BERT, ALBERTA, DISTILBERT, and XLNET** across Machine Learning and Deep Neural Networks.

1.3 Aim and Objectives

The main aim of this research is to study various machine learning and deep learning techniques used for fake review detection and come up novel approach that can be applied using supervised or semi supervised methods for unlabeled data.

• To analyze various Embedding Techniques machine learning and deep learning techniques for fake review detection.
• To propose a suitable feature extraction method and model architecture for fake review detection.
• To assess the effectiveness of different Embedding Techniques such as BERT, RoBERTa, DistilBERT, XLNET and machine learning algorithms, including Support Vector Machine, Naïve-Bayes, and convolutional neural networks (CNNs), recurrent neural networks (RNNs), LSTM, BERT (transformer), Adversarial Learning for automatic feature learning and complex language processing.
• To evaluate the overall performance of the proposed framework using appropriate metrics and validation techniques to assess its efficiency and accuracy in fake review detection.

1.4 Research Questions

Define comprehensive criteria for classifying online reviews into fake reviews and genuine reviews.

- What are the different Embedding Techniques used with ML and Deep Neural Networks models for fake review detection ?
- What are the different ML and Deep Neural Networks techniques used for fake review detection and how are they compared in terms of accuracy, recall, F1 score for fake review detection ?
- How successful are conventional machine learning algorithms like Naive Bayes, Support Vector Machine (SVM), and Logistic regression in identifying fake reviews?
- How have deep learning methods, such as CNN, RNN-LSTM and BERT has been utilized in fake review detection?
- What part do natural language processing Embedding Techniques such as Glove, TF-IDF, BERT embeddings play in these methods?
- To what extent Deep Learning embedding techniques such as BERT, RoBERTa, DistilBERT and XLNet can be utilized for this ?

1.5 Scope of the Study

This study will focus on coming up with a novel approach for Fake review detection by evaluating various machine learning and deep learning techniques. This study aims to come up with novel approach for fake review detection of unlabelled data using semi supervised and unsupervised learning methods.

This study period is expected to be September 2023 till March 2024.

This study will perform literature review of around 40 papers in Fake review detection using machine learning and deep learning techniques,

This study evaluates various machine learning techniques such as word2wec, TF-IDF, BERT embeddings for feature engineering.

This study evaluates various Machine learning methods Logistic regression, SVM, Naïve Bayes and Deep Learning methods such as CNN, CNN-LSTM, RNN-LSTM, BERT.

This study will use data set such as Yelp review data.

1.6 Significance of the Study

This study has significant potential to impact various fields through the accurate detection of fake reviews. Here are some key points:

- **Impact on Various Fields**: Accurate fake review detection can have a profound impact on fields such as social media analysis, opinion mining, and sentiment analysis for product reviews. By identifying deceptive reviews, businesses can gain a more accurate understanding of customer sentiment and improve their products and services accordingly.

- **Understanding Human Deception**: This research can contribute to our understanding of human deception in written language. By analysing the linguistic patterns and strategies used in fake reviews, we can gain insights into how deception is manifested in text. This could have implications for fields such as psychology, communication studies, and even law enforcement.
- **Practical Applications**: The practical applications of this research are vast. In market research, for instance, understanding the sentiment of reviews can help businesses identify market trends and consumer preferences. In the realm of public opinion assessment, detecting fake reviews can ensure a more accurate representation of public sentiment. For customer feedback analysis, it can help businesses distinguish between genuine feedback and spam, enabling them to respond more effectively to their customers' needs.
- **The invention or innovation** in this research has local ramifications for a new body of knowledge, practise, and research in the future.
- **Contribution to the Academic Community:** This thesis research could contribute to the academic community by advancing the understanding of fake review detection and providing valuable resources (e.g., algorithms, datasets, insights) for future research.

1.7 Structure of the Study

The structure of the thesis is as follows.

Chapter 1 section 1.1 presents the background of the research in Fake Review Detection, Section 1.2 discuss the problem statements. Section 1.3 presents aim and objectives of the study. Section 1.4 presents the research contribution to the body of knowledge. The significance of the study provided in section 1.5.

Chapter 2 presents the necessary theoretical background and highlights the problems given in Chapter 1 by systematically reviewing the papers for Fake review detection evolution journey for **Stage 1 Early approaches (Pre-2010)**, **Stage 2 Fake Review Detection using Supervised Learning** and **Stage 3 Fake Review Detection by Unsupervised and Semi Supervised Learning**. This will review highlights of various papers from Pre-2010 till date.

Chapter 3 discusses the **Research Methodology** design and the proposed framework. Section 3.2 describes the research process and introducing the various stages of research methodology such as Data Collection, Data Pre-processing, Model training, and Model Evaluation and Results Analysis ie.

Validation approach to be carried out on the proposed model. **Section 3.3 & Section 3.4** describes Machine learning Algorithms and Deep Learning models.

The summary of the chapter will be given in **Section 3.5**.

Chapter 4 discusses **Implementation** of the research methodology for Fake Review Detection. This covers description of the data set and Exploratory Data Analysis of the Yelp data set, various embedding techniques, implementation of balancing techniques, experimental design.

Chapter 5 discusses **Results and Evaluation.** Here we will discuss the results of various embedding techniques and ML and DL models utilized for fake review detection.

Chapter 6 discusses **Conclusion and recommendations.**

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

A Journey Through Fake Review Detection.

We will study scholarly papers for Fake Review Detection. Imagine a world where online reviews are not genuine, but manufactured praise and engineered discontent. This is how deceptive landscape of fake reviews can be. So getting real truth has become a critical for consumers, businesses, and researchers alike. This paper delves into the evolution of fake review detection, a journey spanning 2005 to 2023 and over 40 scholarly papers, each offering different perspectives.

2.2 Early Approaches (Pre-2010s):

In the early days, pioneers like (Jindal & Liu, 2008) identified three types of review spam, namely i) Deceptive product reviews praising or criticizing unfairly ii) Brand-oriented reviews focusing on manufacturers or sellers, not products. iii) Non-reviews like advertisements or irrelevant text. They suggested outlier reviews must be identified as suspicious review.

(Taboada et al., 2011) delve into a key method for extracting sentiment from text called lexicon-based analysis.

(Mukherjee et al., 2013) In "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews" (2013), explore two main data sets: real reviews from Yelp and artificially generated "pseudo-reviews" created on Amazon Mechanical Turk. The team initially attempts classification using n-gram features, analysing word sequences within reviews. While this approach achieves decent accuracy on pseudo-reviews (67.8%), it performs poorly on real Yelp data. Recognizing this discrepancy, they introduce Kullback-Leibler divergence (KL-divergence), an information-theoretic measure, to pinpoint the specific linguistic differences between genuine and fabricated reviews. KL-divergence analysis reveals fascinating distinctions in writing style. Real reviews demonstrate higher lexical diversity, greater use of adjectives and adverbs, and more subjective expressions. Conversely, pseudo-reviews exhibit repetitive word choices, formulaic sentence structures, and an overreliance on positive sentiment. These findings illuminate the contrasting psychology of genuine reviewers expressing personal opinions and paid workers churning out generic praise. Mukherjee et al. further delve into behavioural features, analysing reviewer activity patterns and engagement. Combining these insights with n-gram analysis significantly improves accuracy, particularly on real-world Yelp data. The study concludes by emphasizing the importance of understanding the distinct characteristics of both real and manufactured reviews to develop effective detection methods. Overall, this paper makes significant contributions to the field of fake review detection. By employing novel analysis techniques and revealing the underlying linguistic and behavioural differences between genuine and inauthentic reviews, Mukherjee et al. provide valuable insights for building more robust and accurate fraud detection systems.

2.3 Machine Learning Approaches

Supervised learning utilizes labeled datasets to train algorithms that accurately classify data or predict outcomes. It requires a corpus of reviews (labelled with real or fake) is typically used for training and testing purposes. We review the supervised learning types of fake review detection by two folds, namely traditional machine learning models and deep learning models.

(Yadav et al., 2021) discusses machine learning techniques such as Naïve Bayes, Random Forest, Logistic Regression and Support Vector Machine (SVM). The papers report that Support Vector Machine (SVM) achieves highest accuracy followed by Radom Forest, Naive Bayes and Logistic Regression.

(Jain et al., 2021) provides Credibility Assessment of User-Generated Content using machine learning techniques on datasets containing Yelp hotel and restaurant reviews. Paper evaluated Logistic Regression (LR), K Nearest Neighbours, Support Vector Machines (SVM), Decision Tree, AdaBoost, Naïve Bayes (NB) and reported LR outperforms with higher accuracy followed by Adaboost, DT, KNN, NB and SVM.

(Al-Zoubi et al., 2023) offers promising solution for combating multilingual spam reviews by the combination of pre-trained language embeddings, optimized machine learning, and language adaptation provides a powerful and adaptable approach. The proposed approach achieves high accuracy in detecting spam reviews for all three languages (English, Spanish, and Arabic). Compared to other methods, WSVM-HHO with BERT embedding yields superior performance across all languages.

(Shunxiang et al., 2023), proposes a model called **SIPUL**, which uses sentiment intensity and **PU** learning to detect fake reviews from streaming data. PU learning is a semi-supervised learning method that only uses positive and unlabelled samples. The assessment of the model indicates that it surpasses the baseline models in terms of accuracy, precision, recall, and F1-score. It demonstrates a strong capability to effectively identify deceptive fake reviews.

(Poonguzhali et al., 2022) proposes using Support Vector Machines (SVM), a powerful machine learning algorithm, to detect fake reviews with high accuracy.

(Zhong et al., 2021) proposing time series analysis to detect sudden spikes in review numbers within specific intervals identified as periods potentially containing busts of fake reviews and machine learning that employs features such as linguistic cues, Metadata and User purchase behaviour.

(Asaad et al., 2023) explores using machine learning algorithms for fake review detection, specifically Support Vector Machines (SVM), Stochastic Gradient Descent (SGD), and XGBoost. It emphasizes analysing review content to identify patterns and characteristics indicative of fakeness.

(Singhal and Kashef, 2023) introduces a novel multi-model architecture, leveraging both the collective intelligence of diverse machine learning algorithms and a targeted sampling strategy, to achieve superior performance in fake review detection. proposes a weighted stacking ensemble model, where individual base learners, including logistic regression and support vector machines, collaborate to form a more robust and accurate classifier. This ensemble's power lies in its heterogeneity and dynamic weighting, which assigns greater influence to base learners demonstrably adept at discerning genuine from deceptive reviews. This approach transcends limitations inherent in singular models, harnessing the combined analytical strength of multiple perspectives.
To further optimize performance, Singhal employs a strategic sampling technique. By focusing on reviews exhibiting characteristics indicative of deception, the model efficiently allocates its analytical resources, prioritizing those most likely to harbour hidden falsehoods. This targeted approach significantly improves detection accuracy while minimizing computational overhead. The empirical results showcase the efficacy of Singhal's framework. The proposed model outperforms baseline approaches, achieving a demonstrably higher degree of precision in identifying fake reviews. This signifies a significant advancement in the fight against online deception, empowering both consumers and businesses to operate with greater confidence in the online marketplace.

2.4 Deep Learning Approaches

Due to the fact that correctly labelled fake review datasets are extremely expensive to create, researchers have developed unsupervised and Semi Supervised machine learning models to identify fake reviews. As fake reviewers improved their skills, the detection race intensified. Deep learning models, fuelled by vast datasets, emerged as the new frontiers.

(Yang et al., 2018) employed **convolutional neural networks** to analyse review structure and identify anomalies, while (Wang et al., 2018) used **long short-term memory recurrent neural network framework** for detecting spanning reviews. These models, though powerful, remained vulnerable to adversarial attacks, where crafty reviewers weaponized their knowledge of the algorithms to game the system (Feng et al., 2020).

(Vaswani et al., 2017)The "Attention is All You Need" paper proposed a groundbreaking neural network architecture called the Transformer for sequence-to-sequence tasks, like machine translation. Unlike previous models relying on recurrent or convolutional layers, the Transformer solely relies on self-attention mechanisms. It introduces the multi-head self-attention mechanism, enabling the model to focus on different parts of the input sequence simultaneously.

(Devlin et al., 2018) The paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" introduced BERT (Bidirectional Encoder Representations from Transformers), a novel language representation model developed by researchers at Google AI. BERT differs from previous models by pre-training a Transformer-based neural network on large amounts of unlabeled text in a bidirectional manner, enabling it to capture context from both left and right contexts in a sentence. This pre-training is conducted through two unsupervised tasks: masked language modeling (MLM) and next sentence prediction (NSP). After pre-training, BERT is fine-tuned on downstream tasks with labeled data, achieving state-of-the-art performance on various natural language understanding tasks, including question answering, text classification, and named entity recognition. The paper demonstrated the effectiveness of BERT across multiple benchmarks and sparked significant advancements in the field of natural language processing.

(Mohawesh et al., 2021) delved into Traditional machine learning methods such as extracting features like review length, vocabulary richness, and sentiment analysis, and utilizing models like Naive Bayes or Support Vector Machines for classification and Deep Larning leverages neural networks, particularly transformers like RoBERTa, to analyse semantic information and context within reviews for improved accuracy. The benchmark study demonstrated that novel deep learning models like RoBERTa can achieve significantly higher accuracy (7% improvement) compared to state-of-the-art methods in a mixed domain. The paper highlights future research areas such as domain-specific models, adapting to evolving spam tactics, and enhancing explainability and transparency.

(Duma et al., 2023) presents a deep hybrid model that merges information from three sources (i) Review text (ii) Overall rating (iii) Aspect Ratings. The model incorporates three components (i) Text Embeddings that capture the semantic meaning of review text using a technique like Word2Vec (ii) Rating Encoders that convert overall and aspect ratings into numerical representations suitable for the model (iii) Deep Fusion Network that combines the extracted features from text, overall rating, and aspect ratings to identify patterns associated with fake reviews. It proposes a bidirectional encoder representation from transformers (BERT) model with a convolutional neural network (CNN) - long short-term memory (LSTM) model to learn hidden text feature vectors.

(Alghamdi et al., 2022), compares the performance of traditional machine learning techniques with hybrid Deep Learning approaches for fake news detection. The conclusion is that Feature Based Models tend to outperform Content Based Model.

.(Deshai & Rao, 2023). Proposes (i) CNN-LSTM model that combines Convolutional Neural Networks (CNNs) for capturing word semantic relationships and Long Short-Term Memory (LSTM) networks for learning long-term dependencies in the review text. (ii) LSTM-RNN model that uses LSTMs in conjunction with Recurrent Neural Networks (RNNs) to analyse the sequential patterns of ratings and identify anomalies suggestive of manipulation. This is used for Fake Rating Detection.

(Deshai and Bhaskara Rao, 2023) proposes two-pronged approach. a) Harnessing power of CNN by scrutinizing both text of the review and additional metadata such as reviewer history and rating patterns. This provides a richer picture for analysing intent and uncovering deception. b) Adaptive PSO (particle swarm optimization)'s intelligent guidance. The Results show that the hybrid system shines with exceptional performance ie 99.4% accuracy and Generalizability across platforms such as tech reviews on Amazon to travel experiences on TripAdvisor, this system adapts to diverse datasets, proving its real-world effectiveness. Deshai acknowledges the ever-evolving nature of fake reviews and suggests further research directions, including utilizing advanced attention mechanisms like Bi-LSTM and BERT for even deeper analysis.

(Zhang et al., 2023) propose two-pronged model ie (i) Behaviour-sensitive feature extractor which analyses how reviewers behave, considering aspects like review frequency, rating patterns, and consistency across different platforms. This helps capture hidden indicators of deception beyond the text itself. (ii) Context-aware attention mechanism that focuses on textual content and uses attention mechanism to identify the most relevant and informative words within each review, allowing the model to better understand the reviewer's intent and sentiment.

(Sasikala et al., 2023), proposes several modifications to the standard CNN architecture, including: (i) Adding max pooling layers which helps capture the most relevant features within each review, improving the model's ability to identify deception. (ii) Incorporating global average pooling which generates a single feature vector representing the entire review, allowing the model to analyse overall sentiment and patterns. (iii) Utilizing a dropout layer that helps prevent overfitting and improve the model's generalizability.

(Ganesh et al., 2023), focuses on two aspects. (i) Benchmarking existing algorithms such as logistic regression, support vector machines, and decision trees, alongside cutting-edge deep learning approaches like bidirectional long short-term memory (BiLSTM). (ii) Proposing novel hybrid methods such as Ensemble of traditional ML algorithms that combines the strengths of multiple traditional algorithms, leveraging their collective learning power to achieve improved accuracy and Hybrid of BiLSTM and metaheuristics: This integrates a BiLSTM network with a metaheuristic optimization technique like evolutionary algorithms. This helps fine-tune the BiLSTM architecture for better feature extraction and classification of fake reviews. Results reveals that Deep learning algorithms generally outperform traditional ML models in terms of accuracy, reaching F1 scores as high as 98%. The proposed hybrid methods achieve state-of-the-art performance, surpassing both individual algorithms and existing ensembles.

(Mewada et al., 2023) proposes SentiBERT, a novel approach that combines sentiment analysis with contextual features to improve fake review detection. SentiBERT uses a pre-trained BERT model to extract contextual features from reviews, and then integrates these features with sentiment features derived from a pre-constructed sentiment dictionary. Finally, a fully connected dense layer classifies the reviews as real or fake using the softmax function. SentiBERT was evaluated on the Yelp dataset and achieved significant improvements over existing methods. It demonstrated a nearly 7% improvement in accuracy compared to existing feature sets and a nearly 4% improvement over existing state-of-the-art methods. SentiBERT offers a promising approach for fake review detection by effectively combining sentiment analysis with contextual features. This can help create a more trustworthy and reliable online environment for consumers and businesses alike.

(Gupta, 2021) discusses the use of transfer learning techniques, specifically **BERT, RoBERTa, ALBERT, and DistilBERT**, for fake review detection in online platforms. The study uses a Yelp dataset comprising over 1.4 million labeled records of fake and genuine reviews. The document outlines the data preprocessing steps, including tokenization and data sampling. Four pre-trained models are implemented and fine-tuned, and their performance is evaluated on 10% and 50% of the dataset.

Results indicate that **RoBERTa**, when trained on 50% of the data, outperforms other models with an accuracy of 69% and a weighted F1-score of 0.69. The study compares the computational time required for training each model and emphasizes the trade-off between accuracy and training time. DistilBERT, despite a 1% lower accuracy, achieves a 43% reduction in training time compared to RoBERTa. Additionally, the document compares the developed models with a baseline SVM model, showing the superiority of RoBERTa.

(Aghakhani et al., 2018) introduces a method called **FakeGAN** for detecting deceptive reviews using **Generative Adversarial Networks (GANs)** and **semi-supervised learning**. The proposed **FakeGAN** leverages two discriminators, D and D', to distinguish between truthful and deceptive reviews. The authors use a dataset of 800 reviews from 20 Chicago hotels, with 400 truthful reviews from TripAdvisor and 400 deceptive reviews written by Amazon Mechanical Workers. **FakeGAN** employs

GloVe vectors for word representation and k-fold cross-validation with k=5 for evaluation. The discriminator accuracy for detecting deceptive reviews reaches approximately 89.2%, comparable to state-of-the-art approaches. The paper discusses the stability, scalability, and performance of FakeGAN, and suggests future work involving different architectures and hyper-tuning. The document concludes by highlighting the potential of FakeGAN for text classification tasks, especially those with large ground truth datasets.

(Salunkhe, 2021) proposes a novel approach for detecting deceptive opinion spam in the paper **Attention-based Bidirectional LSTM for Deceptive Opinion Spam Classification.** The authors employ a Bidirectional Long Short-Term Memory (LSTM) neural network with an attention mechanism. Experimental results demonstrate that the proposed approach outperforms existing methods in terms of accuracy and F1-score, highlighting its effectiveness in identifying deceptive opinion spam.

(Berry and Howard, 2024) explores the phenomenon of fake restaurant reviews on Google and its impact on both consumers and restaurants in the paper **Fake Google restaurant reviews and the implications for consumers and restaurants.** They discuss how fake reviews can distort consumers' perceptions, influence their decision-making process, and harm the reputation of genuine restaurants. Additionally, the paper examines the challenges faced by restaurants in managing their online reputation and suggests strategies for addressing the issue of fake reviews.

2.5 Summary

The paper explores the history and progress of detecting fake reviews from 2005 to 2023, studying over 40 research papers. In the early stages (Pre-2010s), pioneers like Jindal & Liu (2008) categorized review spam into deceptive product reviews, brand-oriented reviews, and non-reviews. Taboada et al. (2011) introduced lexicon-based analysis, while Mukherjee et al. (2013) made significant contributions by using n-gram features and Kullback-Leibler divergence to distinguish real from pseudo-reviews. Their insights into linguistic and behavioral differences between genuine and fake reviews, such as lexical diversity and writing style, laid a foundation for future research.

The paper then delves into supervised learning approaches, showcasing studies like Yadav et al. (2021) favoring SVM and Jain et al. (2021) reporting Logistic Regression's superiority. Additionally, it explores novel solutions for multilingual spam detection, like the WSVM-HHO model by Al-Zoubi et al. (2023). The unsupervised and semi-supervised learning section discusses the vulnerability of deep learning models to adversarial attacks, as highlighted by Feng et al. (2020). The hybrid model proposed by Deshai & Rao (2023), incorporating CNNs, LSTMs, and particle swarm optimization, demonstrates exceptional performance across diverse datasets, emphasizing adaptability and real-world effectiveness. It also touches on transfer learning, using pre-trained models like BERT and RoBERTa, and even explores the application of Generative Adversarial Networks (GANs) with methods like FakeGAN. Overall, the paper provides a comprehensive journey through various detection methodologies, highlighting advancements and challenges in the evolving landscape of fake review detection.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This section will discuss high level stages in research methodology as well as data set description and various machine learning and deep learning techniques that will be employed for fake review detection.

3.2 Research Methodology Framework

3.2.1 Fake Review Detection Methodology Framework

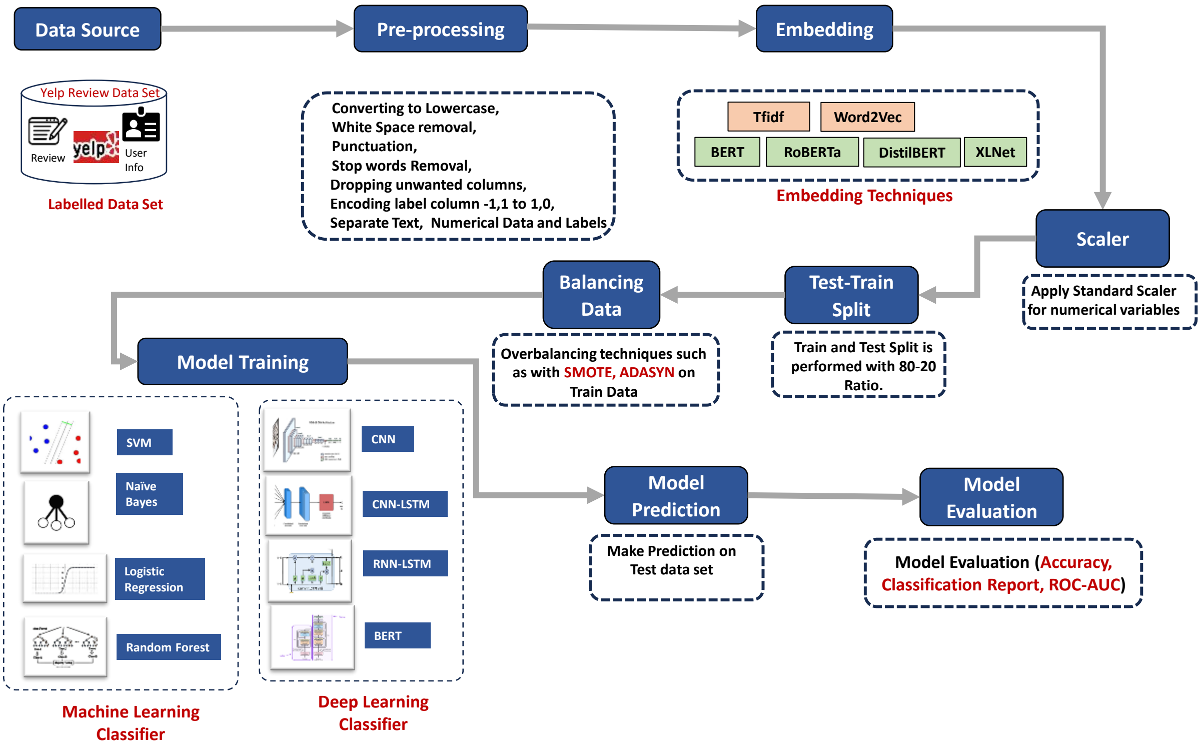


Figure 1 Fake Review Detection Methodology Framework

3.2.2 Proposed Methodology

This study begins with a comprehensive data collection phase, followed by data preprocessing phase, encompassing tasks such as text cleaning, tokenization, and feature extraction to optimize the dataset for model training. Subsequently, traditional machine learning algorithms including Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes, Random Forest (RF) and XGBoost will be implemented. The choice of these algorithms aimed at leveraging their capabilities in handling textual data and discerning patterns indicative of fake reviews.

Following the traditional machine learning phase, the thesis delves into the application of deep learning models. Convolution Neural Networks (CNN), CNN-LSTM, Recurrent Neural Network with Long Short-Term Memory (RNN-LSTM) architecture will be employed to capture sequential dependencies in the reviews, allowing for higher level understanding of context and semantics. Additionally, the study incorporates the use of BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art transformer-based model, renowned for its contextualized word embeddings and advanced language understanding.

Sentence Transformers mentioned under https://sbert.net/, a Python framework for state-of-the-art sentence will be used for text embeddings of fake review detection. Here is the architecture of sentence transformers.

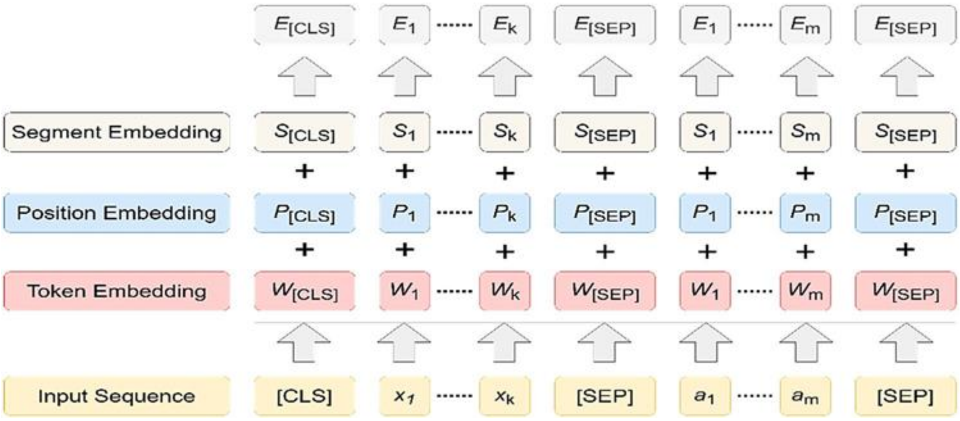


Figure 2 A BERT based Sentence Transformers architecture (Reimers and Gurevych, 2019)

To enhance the adaptability of the models, adaptive learning techniques were integrated into the training process. This adaptive learning approach aimed to optimize the model's performance by dynamically adjusting learning rates based on the observed training progress.

The entire methodology was designed to be iterative, allowing for continuous refinement and optimization of models based on experimental results. The combination of traditional machine learning and advanced deep learning techniques, along with adaptive learning strategies, provided a holistic and sophisticated approach to fake review detection, ensuring a comprehensive exploration of the capabilities of different algorithms and models in handling the intricacies of textual data.

3.2.3 Data Sets:

For our research, the Yelp labelled dataset was utilized. This data set was derived from Yelp Restaurant and Hotel businesses reviews.

Source of data set : https://github.com/yashpandey474/Identification-of-fake-reviews/tree/main/Datasets

3.2.4 Data Pre-Processing:

Data pre-processing is a crucial step in Natural Language Processing (NLP) that involves cleaning and transforming raw text data into a format suitable for analysis and model training. Several techniques are commonly employed in NLP data pre-processing as described below.

Tokenization: This involves breaking down the text into individual words or tokens.

Lowercasing: Converting all text to lowercase helps ensure consistency and reduce the dimensionality of the data.

Stopword removal: Stopwords are common words (e.g., "and," "the," "is") that often don't contribute significant meaning to the text. These words will be removed to reduce the noise and efficiency.

Stemming and Lemmatization: These techniques involve reducing words to their base or root form.

Removing Special Characters and Numbers: Non-alphabetic characters, symbols, and numerical digits might not be essential for some NLP tasks and can be removed.

Handling Contractions and Abbreviations: Expanding contractions (e.g., "don't" to "do not") and standardizing abbreviations help maintain consistency in the language.

Handling Missing Data: Depending on the dataset and task, we might choose to remove, replace, or impute missing values.

Removing HTML Tags and URLs: For web-based text data, removing HTML tags and URLs is essential to extract meaningful content.

Encoding Categorical Variables: For categorical variables (e.g., labels or classes), we might need to encode them numerically for machine learning models to process them effectively.

3.2.4 Embedding Techniques:

Traditional feature engineering may involve creating numerical features based on domain knowledge or extracting statistical properties from raw data. However, in NLP tasks, raw text data cannot be directly fed into machine learning models. Embedding techniques like Word2Vec, GloVe, BERT, and others transform words or text sequences into dense numerical vectors, capturing semantic and contextual information.

TF-IDF: (Term Frequency Inverse document frequency)

Focuses on identifying important words in a document. It considers how often a word appears in the document (frequency) and how rare it is overall (inverse document frequency). This helps prioritize keywords that are specific and essential to the document's content.

GloVe: (Global Vectors for Word Representation)

Captures semantic relationships between words. It analyzes a massive amount of text data to see how often words co-occur, essentially learning that words appearing together frequently share similar meanings. This allows tasks like finding synonyms or analogies based on word vector similarities.

BERT (Bidirectional Encoder Representations from Transformers) :

- A powerful technique for understanding words based on their context in a sentence. It's pre-trained on a huge dataset of text and code, allowing it to learn contextual word representations.
- BERT can be used as an embedding technique to generate contextualized word embeddings for text data. Unlike traditional word embeddings like Word2Vec or GloVe, which produce fixed representations for each word regardless of context, BERT captures contextual information by considering the entire sentence bidirectionally.

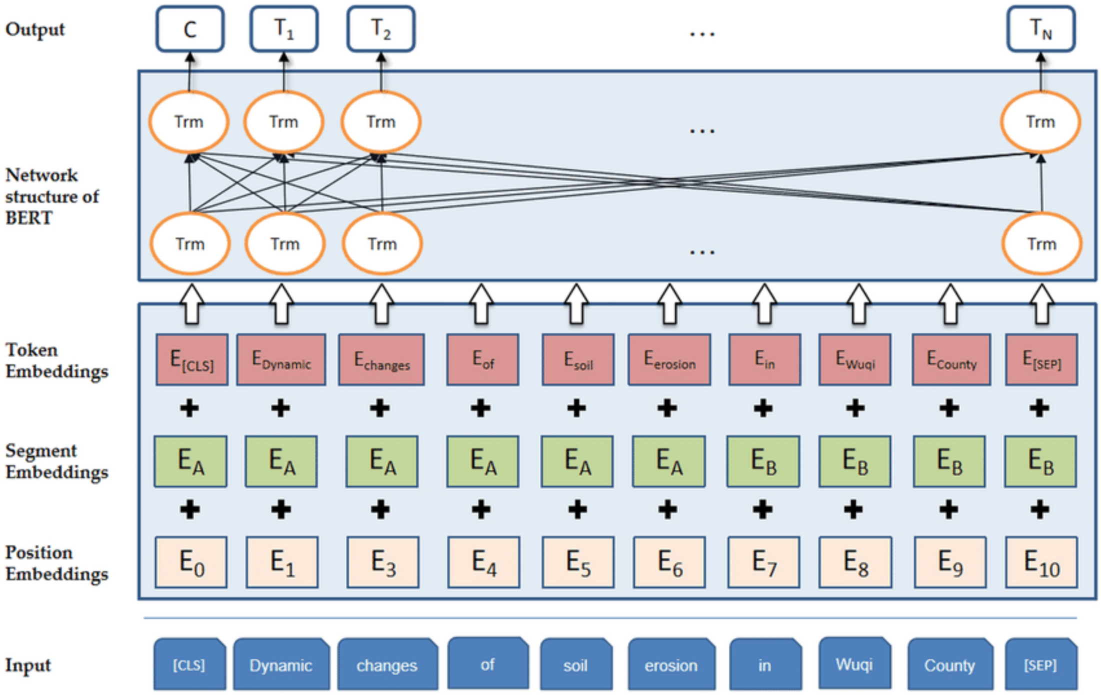


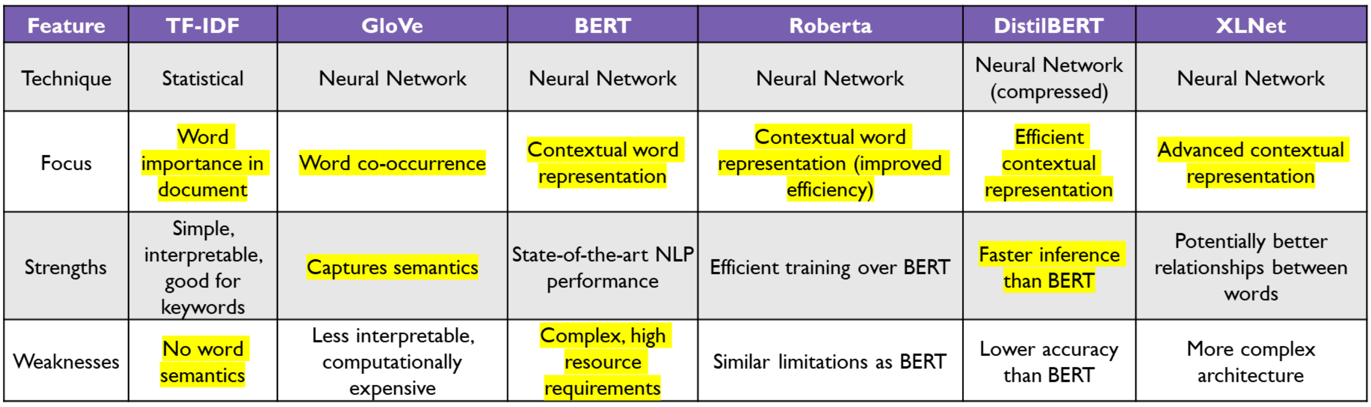
Figure 3 BERT Illustration

Source: https://www.researchgate.net/publication/359301499_Deep_learning-based_methods_for_natural_hazard_named_entity_recognition/figures?lo=1

Roberta: Builds on BERT's success, aiming for better efficiency and performance. It utilizes a more sophisticated masking strategy during training and removes unnecessary steps, making the training process faster and potentially improving performance on certain NLP tasks while retaining BERT's core strengths.

DistilBERT: Creates a compact and efficient version of BERT through a technique called knowledge distillation. It learns from a larger pre-trained model (like BERT) but with a significantly smaller size, allowing for faster processing and lower resource requirements, while maintaining good performance on NLP tasks

XLNet: Addresses limitations in BERT's pre-training process by considering all possible permutations of ordering the input words. This can potentially capture more nuanced relationships between words compared to BERT, but it comes with a more complex architecture and even higher computational demands.



Feature	TF-IDF	GloVe	BERT	Roberta	DistilBERT	XLNet
Technique	Statistical	Neural Network	Neural Network	Neural Network	Neural Network (compressed)	Neural Network
Focus	Word importance in document	Word co-occurrence	Contextual word representation	Contextual word representation (improved efficiency)	Efficient contextual representation	Advanced contextual representation
Strengths	Simple, interpretable, good for keywords	Captures semantics	State-of-the-art NLP performance	Efficient training over BERT	Faster inference than BERT	Potentially better relationships between words
Weaknesses	No word semantics	Less interpretable, computationally expensive	Complex, high resource requirements	Similar limitations as BERT	Lower accuracy than BERT	More complex architecture

3.2.5 Class balancing

Class balancing is crucial in machine learning predictions to overcome the mitigate bias, by balancing the data set, and ensure all classes are treated. It helps prevent the model from favouring the majority class, proper accurate evaluation metrics.

3.2.5.1 SMOTE (Synthetic Minority Over-sampling Technique) is balancing technique that generates synthetic data. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class samples. It works by identifying minority class instances that are close to one another and then creating synthetic instances along the line segments joining them.

3.2.5.2 ADASYN is another balancing techniques which takes care of imbalance by generating synthetic data using K-nearest neighbour of each minority class. Advantage of

ADASYN is that it will not just copy minority data multiple times, but generates more data for complex parts of data which is tricky for ML Algorithm to train on.

3.2.6 Model Training:

Here we train diverse machine learning algorithms and deep learning models on the dataset. First Traditional Machine Learning Algorithms such as **Logistic Regression, Support Vector Machine, Naïve Bayes** will be evaluated for fake review detection. Then Deep Learning Models such as **RNN-LSTM, BERT, Adversarial Learning** will be evaluated for Fake review detection.

3.2.7 Model Evaluation:

Here we will evaluate the performance of the models using appropriate metrics. These will include accuracy, precision, recall, and F1 score. We will also use techniques such as cross-validation to ensure the robustness of our results.

3.2.7.1 Confusion Matrix:

A confusion matrix is a powerful tool in machine learning, particularly for classification tasks. It's a square table that lays out the difference between what your model predicted and what the actual outcome was.

	Predicted Genuine	**Predicted Fake**
Actual Genuine	**TN**	**FP**
Actual Fake	**FN**	**TP**

True Positives (TP): These are the reviews correctly classified as fake by the model.
False Positives (FP): These are the reviews incorrectly classified as fake by the model (i.e., the model predicted them to be fake, but they are actually genuine).
True Negatives (TN): These are the reviews correctly classified as genuine by the model.
False Negatives (FN): These are the reviews incorrectly classified as genuine by the model (i.e., the model predicted them to be genuine, but they are actually fake).

3.2.7.2 Recall (Sensitivity): The proportion of true positives that were correctly identified. Formula:

$$\frac{\mathbf{TP}}{(\mathbf{TP}+\mathbf{FN})}$$

3.2.7.3 Precision: The proportion of true positives among the predicted positives.

Formula :

$$\frac{\mathbf{TP}}{(\mathbf{TP} + \mathbf{FP})}$$

3.2.7.4 Accuracy: The Ratio of proportion of correctly classified instances.

Formula:

$$\frac{(\mathbf{TP} + \mathbf{TN})}{(\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN})}$$

3.2.7.5 F1-score: The F1 score provides a balance between precision and recall, giving a single metric that summarizes the model's performance. It is particularly useful in situations where there is an imbalance between the classes, and both false positives and false negatives need to be considered.

Formula:

$$\frac{\mathbf{2} * (\mathbf{Precision} * \mathbf{Recall})}{(\mathbf{Precision} + \mathbf{Recall})}$$

3.2.7.6 AUC-ROC Curve: Visualizes model performance at various classification thresholds. Area under the curve (AUC) measures model's ability to distinguish classes.

3.2.8 Results Analysis:

In the results analysis of the fake review detection thesis employing machine learning and deep learning techniques, the models will be rigorously evaluated using diverse metrics to gauge their effectiveness.

Performance metrics such as **Accuracy, Precision, Recall,** and **F1 score** will be utilized to assess the classification capabilities of models. A comparative analysis between different models will be done to reveal the variations in their predictive accuracy and computational efficiency, shedding light on the trade-offs inherent in model selection. The comparison with baseline models underscored the advancements achieved in fake review detection through the implemented machine learning and deep learning strategies. The analysis also delves into computational time considerations, addressing the critical balance between model accuracy and training resource requirements. The limitations encountered during experimentation were discussed, contributing to a nuanced interpretation of the results.

Furthermore, future research avenues were suggested, opening possibilities for refining model architectures and exploring additional features to enhance fake review detection systems. In conclusion, the results analysis presented a comprehensive understanding of the strengths, limitations, and potential advancements in the realm of fake review detection using advanced learning techniques.

3.3 Maching Learning Techniques

3.3.1 Logistic Regression

As per (Zou et al., 2019) Logistic Regression serves as a valuable tool for predicting binary outcomes, commonly encountered in diverse domains such as finance, healthcare, manufacturing, and marketing. This method employs a logistic function, also known as the sigmoid function, to model binary dependent variables, often representing categorical distinctions like yes/no or success/failure. The model incorporates a mix of continuous and categorical independent variables, making it versatile for various data types. In a hypothetical scenario with 'n' observations and two features (X1 and X2), Logistic Regression aims to forecast a binary outcome, denoted as Y (0 or 1). Initially, the model computes a linear combination of features, weighted by learned coefficients from training data. Subsequently, it applies the sigmoid function to this calculation to obtain the predicted probability 'p' of the positive class. By selecting a threshold value, typically 0.5, the model assigns binary predictions based on whether 'p' exceeds this threshold. Training involves maximum likelihood estimation to optimize regression coefficients, maximizing the likelihood of observed data. Evaluation metrics such as accuracy, precision, recall, and the F1 score gauge the model's performance. Although Logistic Regression offers interpretability and computational efficiency, it may struggle with non-linear decision boundaries and the assumption of feature independence, potentially limiting its efficacy in complex real-world scenarios.

3.3.2 SVM (Support Vector Machine)

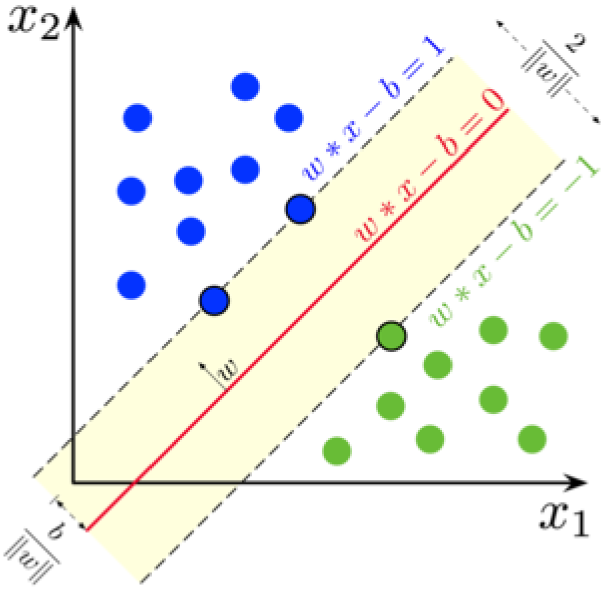


Figure 4 Support Vector Machine

Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

Source : https://en.wikipedia.org/wiki/Support_vector_machine

Support Vector Machine (SVM), a powerful machine learning algorithm, is highly relevant to fake review detection owing to its capability in binary classification tasks. In the context of identifying deceptive reviews, SVM aims to create a hyperplane that effectively separates the feature space into distinct regions corresponding to genuine and fake reviews. The algorithm achieves this by maximizing the margin between the two classes, selecting the hyperplane that optimally segregates them. SVM is particularly useful when dealing with complex, non-linear relationships between features, as it can employ kernel functions to map the input data into a higher-dimensional space. This allows SVM to capture intricate patterns and make nuanced distinctions between authentic and fake reviews. The versatility of SVM, along with its ability to handle high-dimensional data, makes it a valuable tool for the nuanced and challenging task of fake review detection.

3.3.3 Naïve Bayes

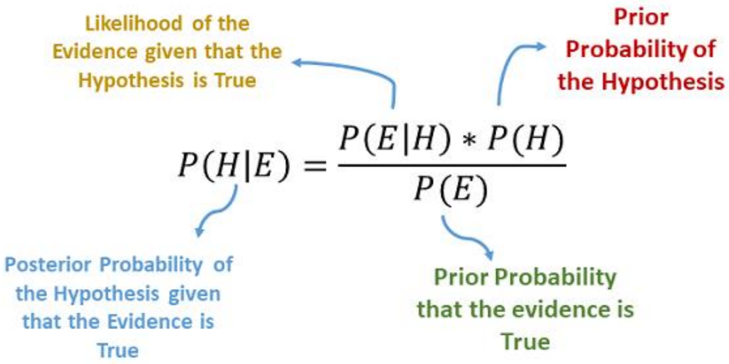

Likelihood of the Evidence given that the Hypothesis is True
Prior Probability of the Hypothesis
$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$
Posterior Probability of the Hypothesis given that the Evidence is True
Prior Probability that the evidence is True

Figure 5 Naive Bayes

Source : https://medium.com/analytics-vidhya/na%C3%AFve-bayes-algorithm-5bf31e9032a2

The **Naive Bayes algorithm** is highly relevant to fake review detection, offering a probabilistic approach to classification tasks. In the context of distinguishing between genuine and deceptive reviews, Naive Bayes relies on Bayes' theorem and assumes that the features used for classification are conditionally independent. Despite its "naive" assumption, the algorithm is effective in practice, especially when dealing with text data. In fake review detection, Naive Bayes calculates the probability that a given set of features (words or phrases) belongs to a specific class, allowing it to assess the likelihood of a review being genuine or fake based on observed word occurrences. Its simplicity, speed, and efficiency in handling high-dimensional data, such as the language used in reviews, make Naive Bayes a popular choice for text classification tasks, including the challenging domain of fake review detection.

3.3.5 Random Forest

(Zou et al., 2019) Random Forest is a potent supervised learning technique extensively utilized for tasks like classification and regression, including the detection of fake reviews. This algorithm operates by constructing numerous decision trees during training, where the final output is determined by the mode of classes for classification or the mean prediction for regression from individual trees.

In essence, a Random Forest comprises a multitude of independent decision trees. These trees utilize the Gini index, a metric of inequality, to select splitting attributes. The depth of each tree branch depends on the algorithm's parameter 'd'.

The algorithm creates several decision trees, each trained on a random subset of the training data and a random subset of features. This randomness in data and feature selection for each tree helps prevent overfitting, thus enhancing model accuracy. Following training on these different subsets, the algorithm aggregates the predictions from all trees to make a final decision. This process also facilitates feature importance ranking and simplifies the learning process to optimize model computation.

Although Random Forest lacks a straightforward mathematical formula like Logistic Regression or a Decision Tree due to its ensemble nature, its operational principle can be represented mathematically.

In a Random Forest with 'n' decision trees, each tree produces a classification outcome denoted as Yj, where 'j' is the tree index. For binary classification, each Yj can be either 0 or 1.

The Random Forest classifier's output, Y_RF, is determined by the majority vote (for classification) or the average (for regression) of the individual decision trees' outputs. For classification, this can be expressed mathematically as:

Y_RF = mode(Y1, Y2, ..., Yn)

For regression, the output is the average of the individual trees' outputs:

Y_RF = (Y1 + Y2 + ... + Yn) / n

Each decision tree is constructed with a bootstrapped sample of the original data, a technique known as bagging or bootstrap aggregating. Additionally, at each candidate split during learning, a random subset of features is considered, contributing to the model's robustness and control overfitting.

While Random Forest offers flexibility and can model complex feature interactions effectively, it may underperform compared to simpler models in scenarios where relationships are straightforward and linear. Moreover, it requires more computational resources and memory. In summary, Random Forest stands as a versatile ensemble learning method employing multiple decision trees to generate final predictions, thus serving as a robust tool for various machine learning tasks, including the detection of fake reviews.

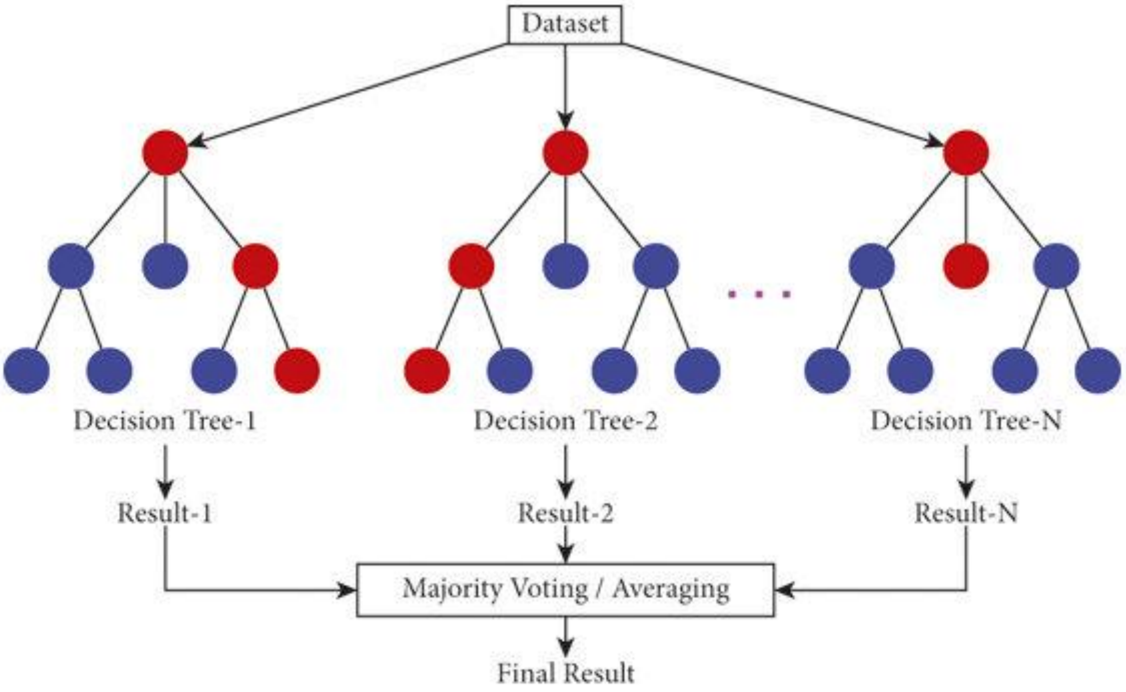


Figure 6 Random Forest Illustration

Source : https://www.researchgate.net/publication/354354484_Automated_Prediction_of_Good_Dictionary_EXamples_GDEX_A_Comprehensive_Experiment_with_Distant_Supervision_Machine_Learning_and_Word_Embedding-Based_Deep_Learning_Techniques/figures?lo=1

3.3.6 XGBoost

XGBoost, or eXtreme Gradient Boosting, is a high-level machine learning method within a gradient boosting framework. It's a decision tree-based approach to improve the speed and accuracy of traditional decision trees. XGBoost is an optimized implementation of gradient boosting, commonly used for regression and classification, forming a model ensemble from weak prediction models, usually decision trees.

XGBoost builds upon gradient boosting, incorporating additional regularization to prevent overfitting and enhance overall model performance. The objective function optimized by XGBoost for binary labeled instances can be expressed as:

$$Obj(\Theta) = \sum l(yi, \hat{y}i) + \sum \Omega(f\_k)$$

Here, $l(yi, \hat{y}i)$ measures the difference between the actual target yi and the predicted outcome $\hat{y}\_i$, while $\Omega(fk)$ penalizes model complexity by considering factors like the number of leaves in the tree and the L2 norm of leaf scores.

The regularization term $\Omega(fk)$ is defined as:

$$\Omega(fk) = \gamma T + 1/2\ \lambda\ ||w||\string^2$$

Where T signifies the number of leaves, w represents leaf scores, and γ and λ are regularization parameters.

XGBoost optimizes this objective function using gradient boosting, a model that improves prediction accuracy by repeatedly adjusting its parameters to reduce errors. Each iteration constructs a new model, usually a decision tree, to correct errors made by the preceding model. This process continues until the error is minimized.

While this summary outlines the core mathematics behind XGBoost, its implementation involves complex procedures such as computing optimal leaf weights, pruning trees, and handling missing data, which are beyond this explanation.

XGBoost offers several advantages over traditional methods, including built-in regularization, handling missing values, and efficient processing, making it a powerful and versatile machine learning algorithm for various tasks.

3.4 Deep Learning Models.

3.4.1 Convolution Neural Networks (CNN)

Convolutional Neural Networks (CNNs) have proven to be effective in fake review detection by capturing hierarchical patterns and local dependencies within textual data. In the context of fake review identification, CNNs operate by employing a series of convolutional filters over the input text, allowing the model to automatically learn relevant features at different levels of abstraction. These filters, or kernels, convolve across the input review, detecting patterns like specific linguistic structures or deceptive language usage. The subsequent pooling layers help extract the most salient information from the convolved features. By leveraging these hierarchical features, CNNs can discern subtle textual nuances indicative of fake reviews, making them adept at capturing both local and global contextual information essential for accurate classification in the realm of deceptive online reviews.

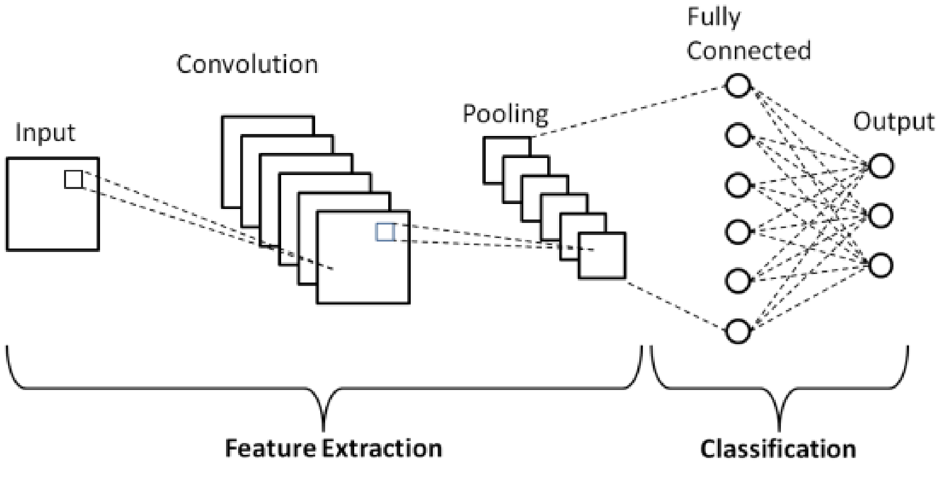


Figure 7 CNN architecture

Source - https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53

3.4.2 Recurrent Neural Networks (RNN)

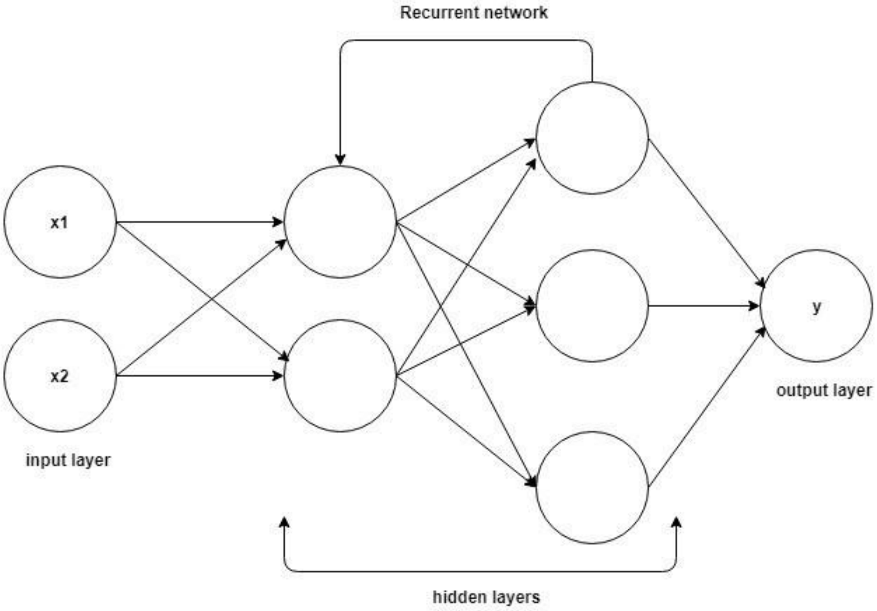


Figure 8 RNN Architecture

Source - https://towardsdatascience.com/machine-learning-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebc

Recurrent Neural Networks (RNNs) are pertinent to fake review detection, particularly in handling sequential data like textual reviews. RNNs are designed to capture dependencies and patterns over time, making them well-suited for tasks where the order of information matters. In the context of identifying fake reviews, RNNs can analyse the sequential nature of language, considering the context and relationships between words in a review. By maintaining a hidden state that evolves as new words are processed, RNNs can capture long-range dependencies that may be crucial for discerning deceptive language. However, traditional RNNs face challenges in capturing long-term dependencies due to vanishing or exploding gradient problems. More advanced variants like Long Short-Term Memory (LSTM) networks have been introduced to address these issues, enhancing the effectiveness of RNNs in modeling sequential data and improving their applicability to tasks such as fake review detection.

3.4.3 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are highly relevant to fake review detection, especially when dealing with the complexities of sequential data. LSTMs belong to the family of Recurrent Neural Networks (RNNs) and are designed to address the vanishing and exploding gradient problems encountered by traditional RNNs. In the context of identifying fake reviews, LSTMs excel at capturing long-term dependencies and relationships between words in a review. The architecture of LSTMs includes memory cells and gating mechanisms that allow them to selectively remember or forget information over extended sequences, making them particularly effective for tasks that involve analysing the sequential structure of language. This capability is crucial for detecting subtle linguistic patterns and understanding the context in which certain phrases or expressions are used, enhancing the accuracy of fake review detection models.

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to overcome the limitations of traditional RNNs in capturing long-term dependencies in sequential data. LSTMs were introduced to address the vanishing gradient problem, which hinders the training of RNNs on sequences with long-term dependencies.

In an LSTM network, the key components are memory cells, input gates, forget gates, and output gates. Memory cells allow the network to maintain a long-term memory, and the gates regulate the flow of information into, out of, and within the cell. The input gate controls the information to be stored in the memory cell, the forget gate decides what information to discard from the cell, and the output gate determines the output based on the current input and the internal memory.

The architecture's ability to selectively remember or forget information allows LSTMs to capture relevant patterns over extended sequences, making them effective for tasks involving sequential data, such as natural language processing, speech recognition, and time-series prediction. The sophisticated gating mechanism enables LSTMs to mitigate the vanishing gradient problem, making them more capable of learning and retaining information over long-range dependencies in sequential input.

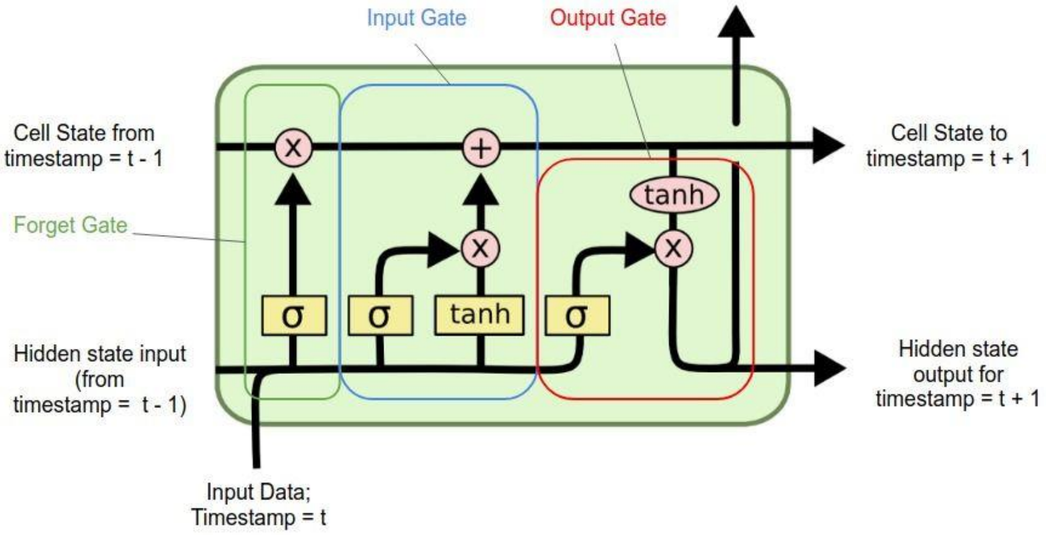

Input Gate
Output Gate
Cell State from timestamp = t - 1
Forget Gate
Hidden state input (from timestamp = t - 1)
Input Data; Timestamp = t
x
+
σ
σ
tanh
σ
tanh
x
x
Cell State to timestamp = t + 1
Hidden state output for timestamp = t + 1

Figure 9 LSTM Architecture

Source - https://www.turing.com/kb/comprehensive-guide-to-lstm-rnn

3.4.4 RNN-LSTM

Combining **Recurrent Neural Networks (RNNs)** with **Long Short-Term Memory (LSTM)** cells, known as **RNN-LSTM**, is particularly relevant in the context of fake review detection. RNN-LSTM models are well-suited for processing sequential data, such as text, by addressing the challenges of capturing long-term dependencies. In the realm of fake review detection, RNN-LSTM can effectively analyse reviews over an extended context, capturing nuanced linguistic patterns and dependencies between words. The recurrent connections in RNNs allow information to persist over time, while the LSTM cells mitigate the vanishing and exploding gradient problems associated with traditional RNNs. This combination enables the model to better understand the sequential nature of language and identify subtle cues indicative of fake reviews, making RNN-LSTM a powerful tool for enhancing the accuracy of detection algorithms in this domain.

3.4.5 BERT (Bidirectional Encoder Representations from Transformers)

BERT (Bidirectional Encoder Representations from Transformers) is highly relevant to fake review detection due to its exceptional ability to comprehend contextual nuances in language. BERT, based on the Transformer architecture, processes words bidirectionally, considering the entire context of a word within a sentence. In the context of fake review detection, BERT excels in capturing intricate relationships between words, discerning sarcasm, and understanding the subtle contextual cues that may indicate deceptive reviews. Pre-trained on massive corpora, BERT brings a wealth of contextual knowledge to the task. Fine-tuning BERT for fake review detection involves training it on labeled datasets, allowing it to adapt its contextual understanding specifically for distinguishing between genuine and fake reviews. The contextual embeddings produced by BERT provide a robust foundation for building accurate and nuanced classifiers, making it a potent tool in the arsenal of techniques aimed at identifying deceptive reviews..

3.5 Summary

The research methodology outlined in the provided text involves a systematic approach to detect fake reviews using a combination of traditional machine learning and deep learning techniques. Here's a summary:

3.5.1. Introduction:

- Overview of high-level research methodology stages.
- Use of the Yelp dataset, a publicly available dataset, for investigation.

3.5.2. Methodology Description:

- Block Diagram:

- Introduction of the Fake Review Detection Methodology Framework.

- **Proposed Methodology:**
  - **Data Collection:**
    - Utilization of the Yelp dataset containing information on businesses, reviews, and users.
  - **Data Pre-Processing:**
    - Application of Natural Language Processing (NLP) techniques:
      - Tokenization, lowercasing, stopword removal, stemming/lemmatization.
      - Special character/number removal, handling contractions/abbreviations.
      - Handling missing data, removing HTML/URLs.
      - Encoding categorical variables, text vectorization.
  - **Model Training:**
    - Training diverse models:
      - Traditional Machine Learning (Logistic Regression, SVM, Naïve Bayes).
      - Deep Learning Models (RNN-LSTM, BERT).
  - **Model Evaluation:**
    - Performance evaluation using metrics such as Accuracy, Precision, Recall, F1 score.
    - Confusion matrix and AUC-ROC curve used for comprehensive analysis.
    - Cross-validation employed for result robustness.
  - **Results Analysis:**
    - Rigorous evaluation of models and comparative analysis using metrics such as accuracy, precision, recall, F1 score.
    - Consideration of computational time and discussion of limitations.
    - Suggestions for future research and model improvements.

3.5.3. Machine Learning Algorithms Description:

- Overview of key machine learning algorithms:
  - Logistic Regression, SVM, Naïve Bayes, Random Forest, XGBoost

3.5.4. Deep Learning Models Description:

- Overview of deep learning models:
  - CNN, RNN, LSTM, RNN-LSTM, BERT.
  - Each model's relevance to fake review detection explained.

The methodology combines traditional machine learning and advanced deep learning techniques to comprehensively address the challenges of detecting fake reviews. The iterative nature of the methodology allows for continuous refinement based on experimental results, providing a holistic exploration of various algorithms and models in handling textual data intricacies.

CHAPTER 4: IMPLEMENTATION

4.1 Introduction

This section covers various techniques used in implementation of Methodology for Fake Review Detection. As part of this implementation a comprehensive **Exploratory Data Analysis (EDA)** is conducted, serving as a cornerstone of our research.

Exploratory Data Analysis (EDA) in the context of **Natural Language Processing (NLP)** and **Fake review detection** involves analyzing and understanding the textual data in the dataset. Here are some key steps and techniques for performing EDA specific to NLP.

4.2 Exploratory Data Analysis (EDA)

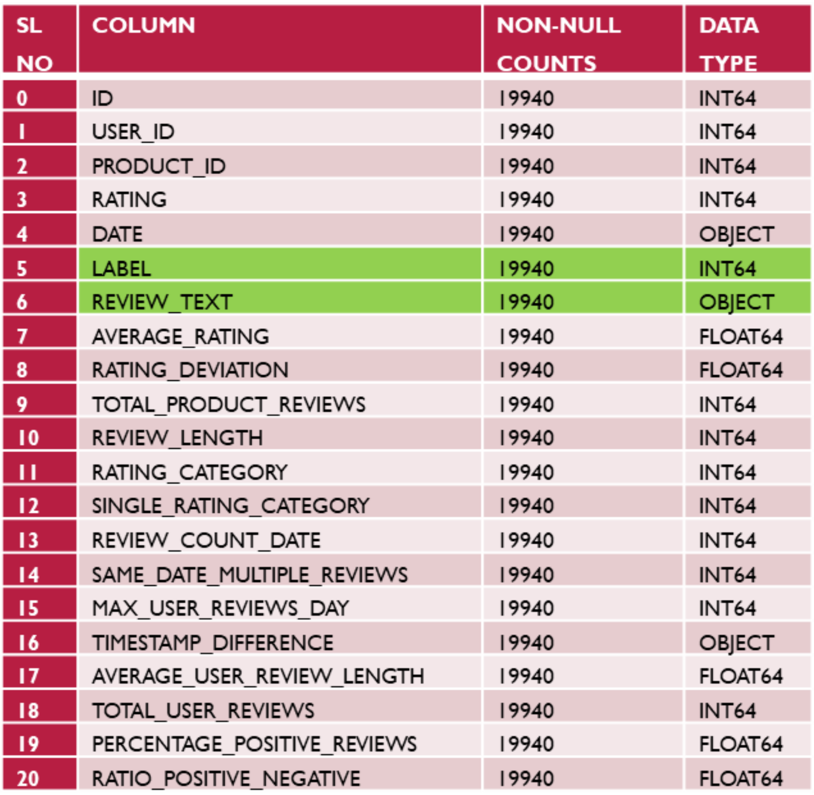
4.2.1 Data Set Description

Yelp data is used for the study of fake review detection. This is a labeled data set and derived from Yelp data set based on Yelp data available in public domain.

Source of data set : https://github.com/yashpandey474/Identification-of-fake-reviews/tree/main/Datasets

4.2.2 Data definition of Yelp Labeled Data Set.

Table 1Yelp Labelled Data Set Description



SL NO	COLUMN	NON-NULL COUNTS	DATA TYPE
0	ID	19940	INT64
1	USER_ID	19940	INT64
2	PRODUCT_ID	19940	INT64
3	RATING	19940	INT64
4	DATE	19940	OBJECT
5	LABEL	19940	INT64
6	REVIEW_TEXT	19940	OBJECT
7	AVERAGE_RATING	19940	FLOAT64
8	RATING_DEVIATION	19940	FLOAT64
9	TOTAL_PRODUCT_REVIEWS	19940	INT64
10	REVIEW_LENGTH	19940	INT64
11	RATING_CATEGORY	19940	INT64
12	SINGLE_RATING_CATEGORY	19940	INT64
13	REVIEW_COUNT_DATE	19940	INT64
14	SAME_DATE_MULTIPLE_REVIEWS	19940	INT64
15	MAX_USER_REVIEWS_DAY	19940	INT64
16	TIMESTAMP_DIFFERENCE	19940	OBJECT
17	AVERAGE_USER_REVIEW_LENGTH	19940	FLOAT64
18	TOTAL_USER_REVIEWS	19940	INT64
19	PERCENTAGE_POSITIVE_REVIEWS	19940	FLOAT64
20	RATIO_POSITIVE_NEGATIVE	19940	FLOAT64

This data set had 19940 rows.

This data set has 21 columns.

Out of 21 columns, 2 columns will be utilized for this study. These columns are **REVIEW_TEXT and LABEL.**

4.2.3 Distribution of target variable ie LABEL (Fake vs. Genuine)

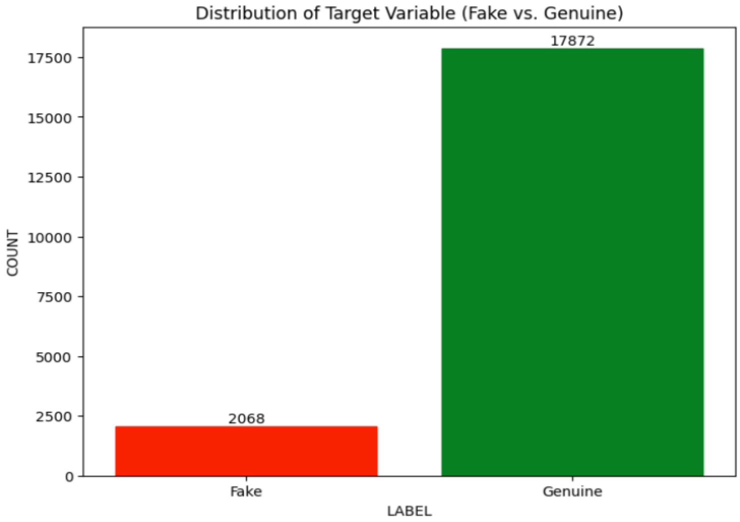


Figure 10 Distribution of target variable ie Label

Out of 19940 rows, 17872 are Genuine reviews and 2068 are Fake Reviews.

4.2.5 Distribution of Review Text lengths

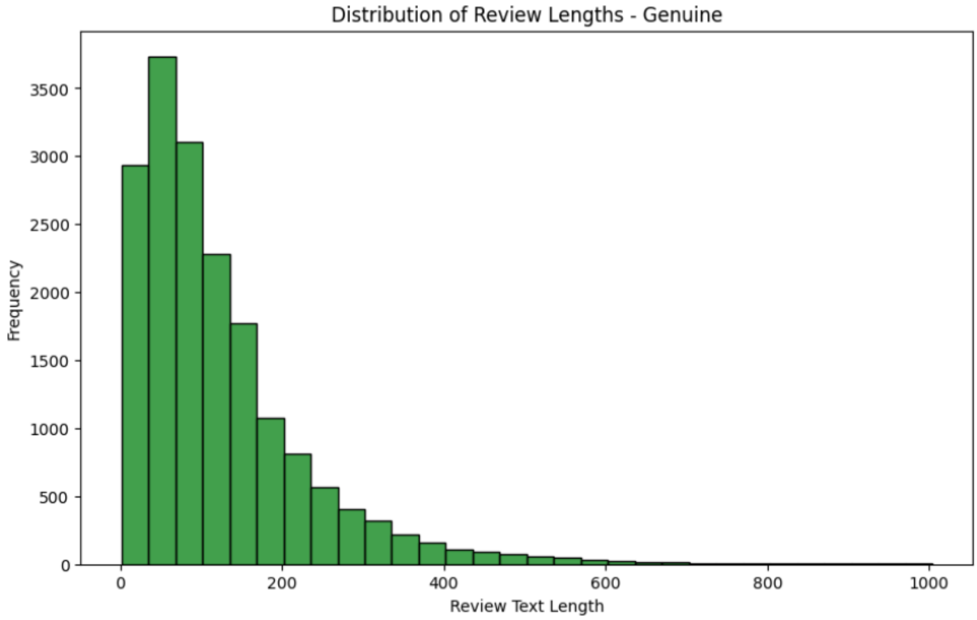


Figure 11 Distribution of Genuine Review Text Lenghts

Distribution of data lengths show that this

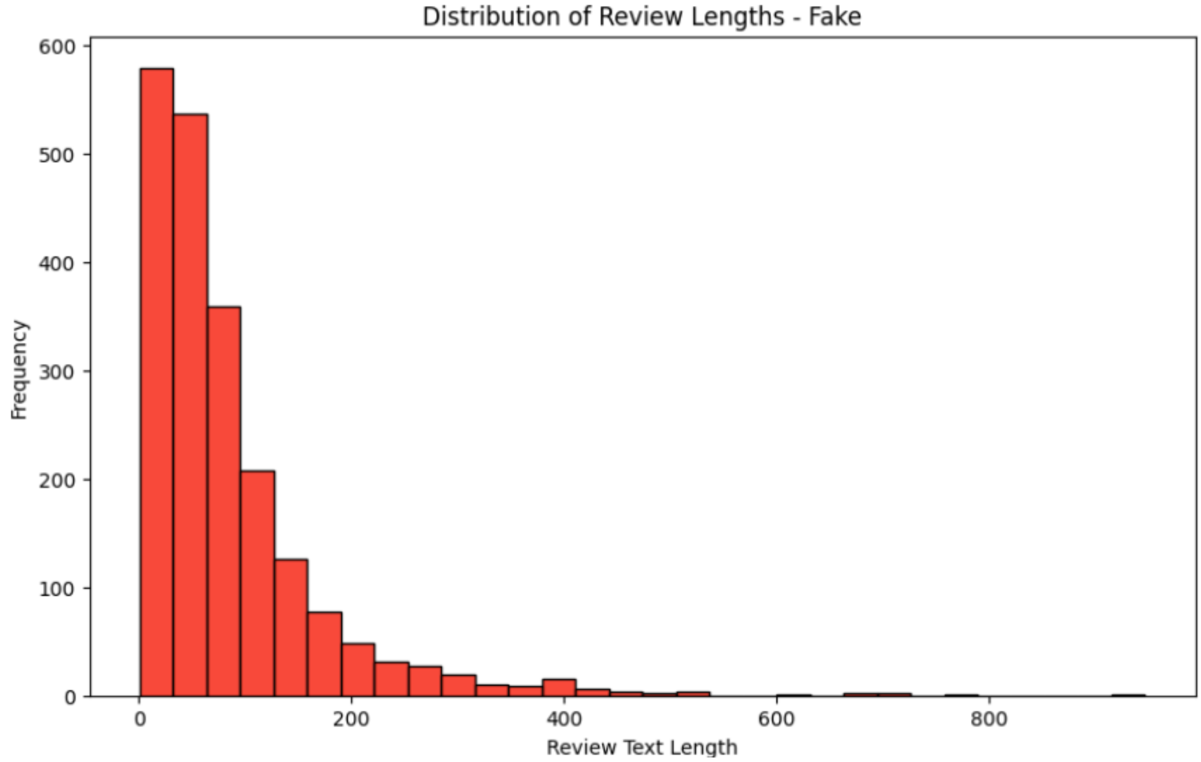


Figure 12 Distribution of Fake Review Text Lenghts

4.2.6 Check the dimensions of the dataset

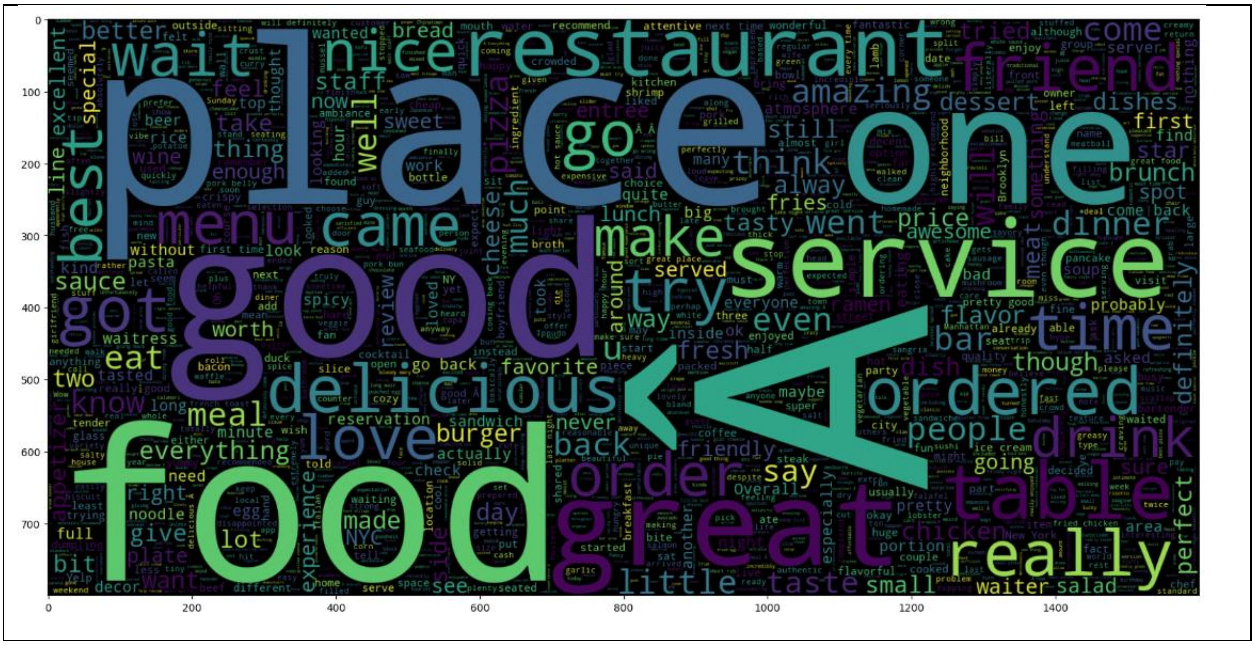
Dataset Dimensions:
Number of rows: 19940
Number of columns: 22

4.2.7 Sample Fake reviews vs Genuine Reviews

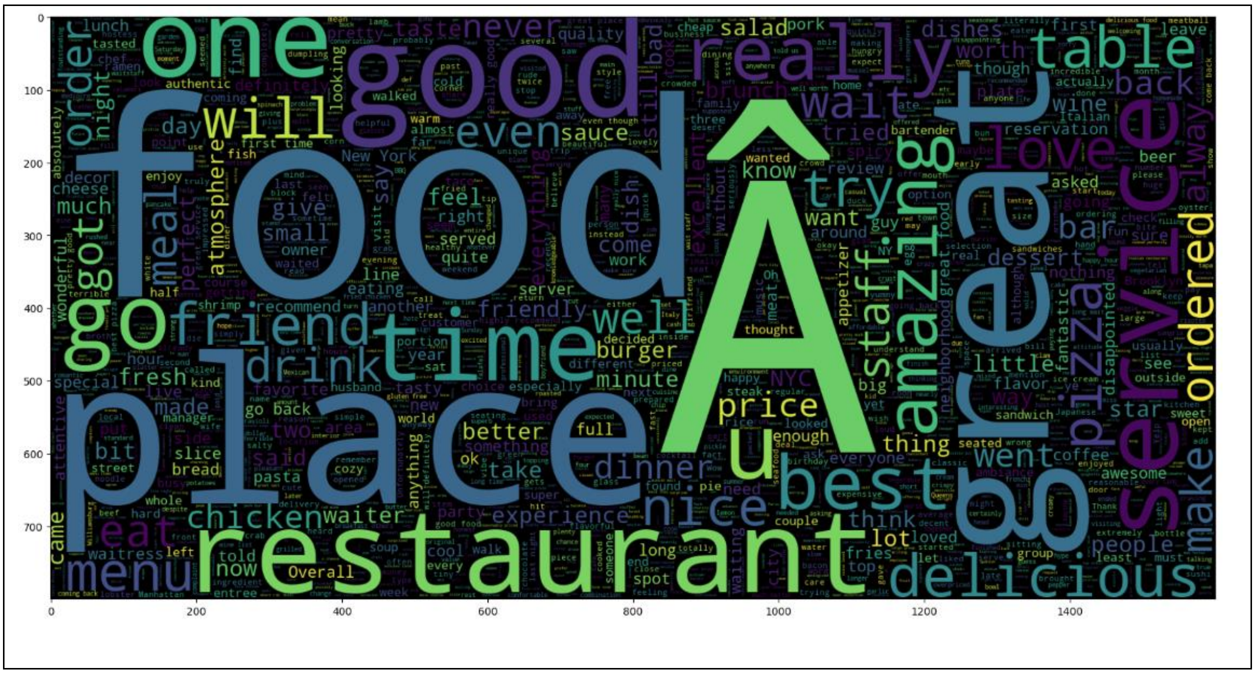
Fake Review	Genuine Review
0 Great.....	1 My family and I had Bubby's brunch on a Saturday morning. Â We got there at 10am and there wasn't much of a wait. Â Tribeca must be a family friendly neighborhood since everyone parked their strollers outside without worry. Â For those with itty bitty ones they do have high-chairs and the changing table is located downstairs. Â As for the food, I had the Bubby's Breakfast with homefries and ham. Â The ham was a bit dry but everything else on the plate was good. Â I had a taste of my sister's blueberry pancakes and it was delicious. Â The only negative was what looked like a chunk of cheese on the lip of my coffee cup. Â I don't think they make French Onion coffee. Â Our waitress apologized and replaced the cup. Â At $30 per person it's expensive for breakfast + coffee + tax but this isn't your neighborhood IHOP.
35 Great halal food!	2 I really like this place, but they need to get their menu sorted out. I finally went for brunch, which I'd been eager to try based on reviews and the mouth-watering online menu (sugarfreak.com/fooddrinkâ€¦), but was disheartened when the baked apple with moonshine double cream wasn't available. That's just cruel!
100 Spectacular food, great service, and an amazing open service area. I like the prix fixe menu. The food was just amazing, spiced perfectly and presented beautifully. While the main courses are small, I certainly did not leave hungry. The wine was excellent and I am pining to go back already!	3 This is one of my favorite places in the US. Awesome! No reservations and the place was packed but the maÃ®tre d' squeezed us in when someone didn't arrive on time. Carol was our server and she was really helpful and made some great recommendations. Simply the best steak tartare!
103 We went for happy hour and shared the swordfish ceviche on plantain crackers and 1/2 chicken. Â Both dishes were delicious and almost too much for two. The service was so friendly and attentive, the bar tender answering our many questions cheerfully. Â We're going to go back for a late night band session and cocktails now!	4 Make sure you go with a small group of friends that ALL like to share food so you can try 1 of everything that everyone on here is talking about. Â I loved it ALL except I won't know about the breakfast sandwich until the next time I go since this one girl at our table wouldn't give me a bite. Â C'mon! Â I'm always a Savory over Sweet breakfast lover BUT I had a foodgasm for the (you guessed it) blueberry pancakes. Â I'm thinking about the blueberry pancakes right now......

4.2.8 Word Cloud Fake Reviews vs Genuine Reviws

Word Cloud Fake Reviews



Word Cloud Genuine Reviws



4.3 Resource Requirements

4.3.1 Hardware Requirements

A computer with high processing power at least 16GB of RAM (32GB preferable), quad-core processor, 500GB of storage space

4.3.2 Software Requirement

- Python programming language version 3.8 or higher.
- TensorFlow version 2.4 or higher.
- Scikit-learn version 1.0.3 or higher.
- Keras version 2.4 or higher.
- Matplotlib version 3.7.1 or higher.
- Seaborn version 0.11.2 or higher
- Pandas
- Numpy
- Jupyter Notebook version 6.1 or higher, an interactive computing environment for Python.
- A version control system like Github to track and manage changes to the code and data throughout the research process.
- Imblearn – Imblearn.Pipeline - Used for building pipeline to merge text and numerical data, scaling, sampling technique.
- Imblearn.Smote used for balancing of the data.
- SVC, MultinomialNB, LogisticRegression, RandomForestClassifier from Scikit-learn library for Python.
- Transformers
    - Bert pre-trained model - bert-base-uncased tokenization
    - Roberta pre-trained model - roberta-base tokenization
    - DistilBERT pre-trained model - distilbert-base-uncased tokenization
    - XLNET pre-trained model - xlnet-base-cased tokenization

CHAPTER 5: RESULTS AND EVALUATION

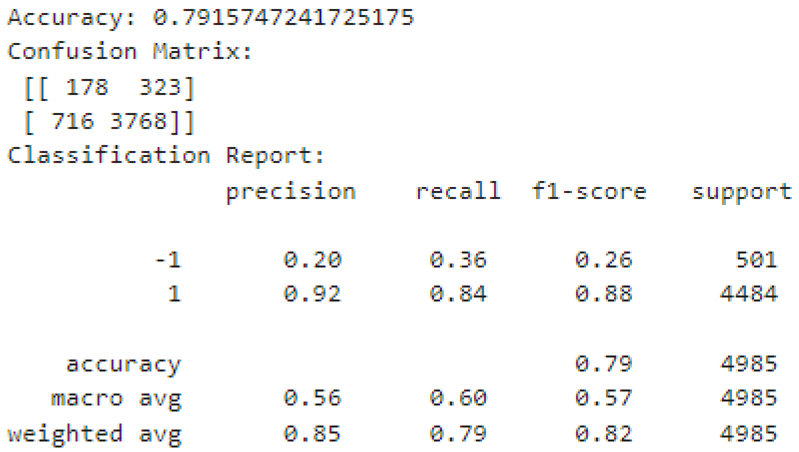
5.1 Introduction

This section discusses how **Machine Learning algorithms**, namely **Logistic Regression, Support Vector Machine, Random Forest, XGBoost** and **Deep Learning Models** such as **CNN (Convolution Neural Network), CNN-LSTM, RNN -LSTM, BERT (Bidirectional Encoder Representations from Transformers)** applied for the fake review detection. During this process various embedding techniques such as TFIDF and BERT, ROBERTA, DISTEILBERT, XLNET embeddings were explored. Balancing techniques such as ADASYN, SMOTE (Oversampling minority classes) were tested. Standard Scaler were applied for numerical columns. It noted that Deep Learning Models MLP Classifier with BERT, ROBERTA, DISTEILBERT, XLNET Embeddings outperformed other Embedding techniques TFIDF and Glove on Machine Learning Classifiers.

5.2.1 TFIDF Embedding Comparison with Machine Learning Classifiers

Here we will discuss the results of various metrics captured for each ML Models.

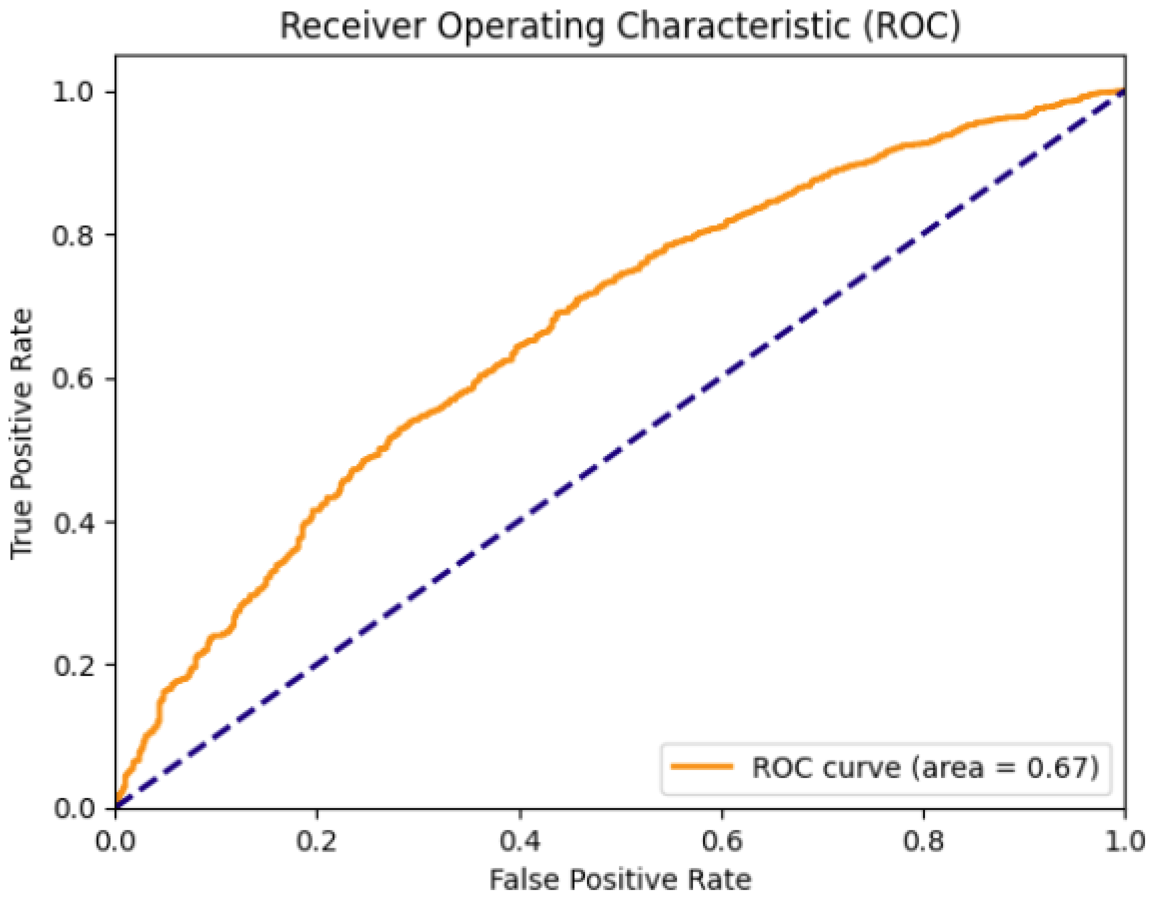
5.2.1.1 Logistic Regression



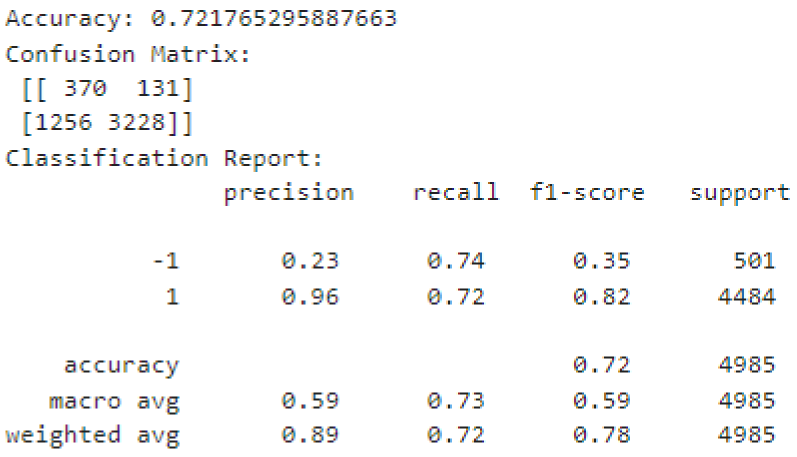
```
Accuracy: 0.7915747241725175
Confusion Matrix:
 [[ 178  323]
 [ 716 3768]]
Classification Report:
              precision    recall  f1-score   support

          -1       0.20      0.36      0.26       501
           1       0.92      0.84      0.88      4484

    accuracy                           0.79      4985
   macro avg       0.56      0.60      0.57      4985
weighted avg       0.85      0.79      0.82      4985
```



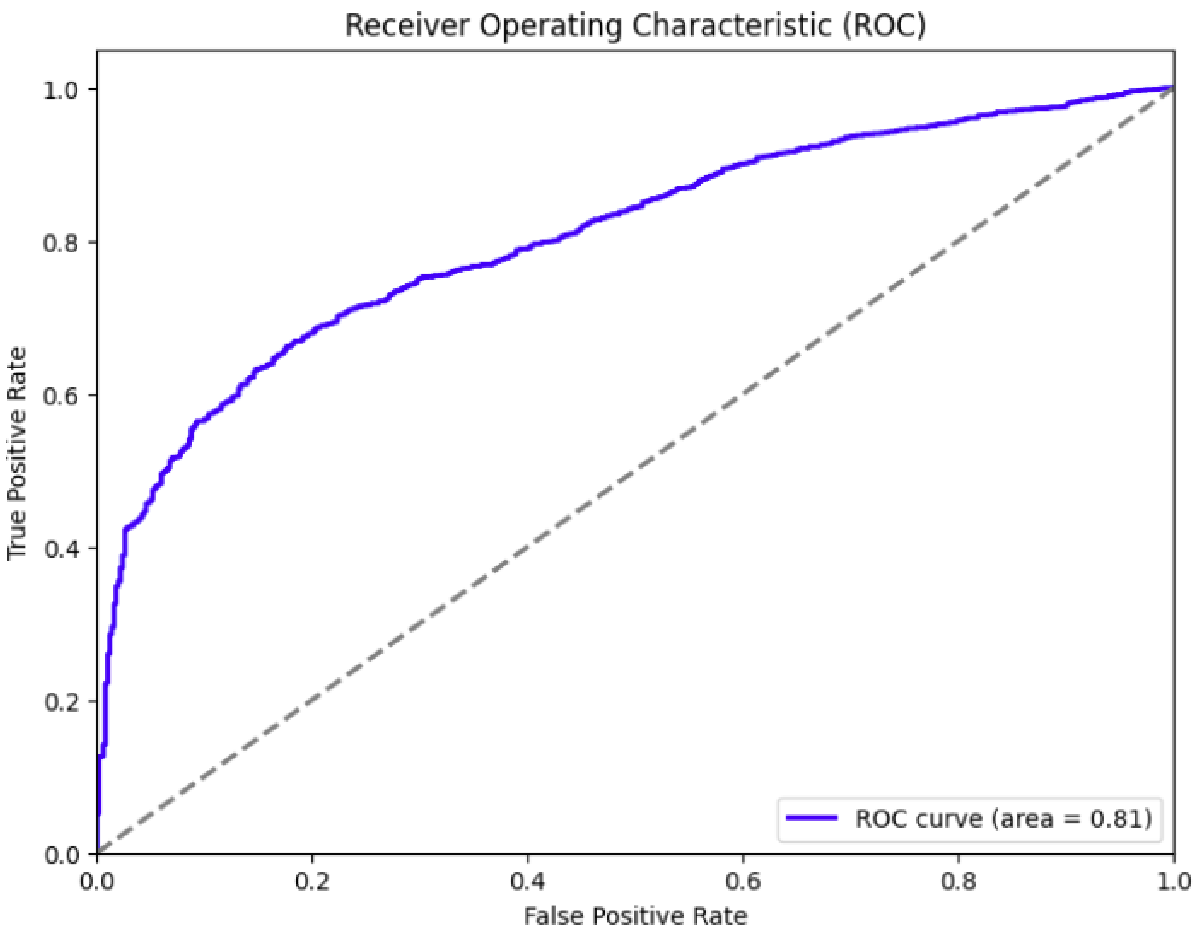
5.2.2.2 Support Vector Machine



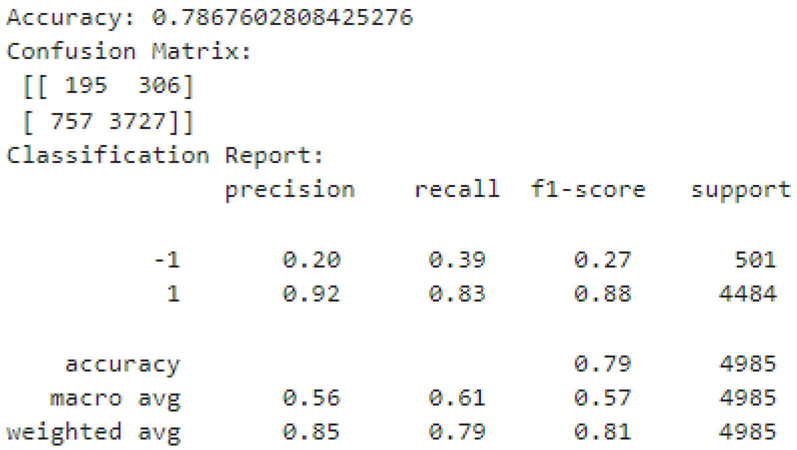
```
Accuracy: 0.721765295887663
Confusion Matrix:
 [[ 370  131]
 [1256 3228]]
Classification Report:
              precision    recall  f1-score   support

          -1       0.23      0.74      0.35       501
           1       0.96      0.72      0.82      4484

    accuracy                           0.72      4985
   macro avg       0.59      0.73      0.59      4985
weighted avg       0.89      0.72      0.78      4985
```



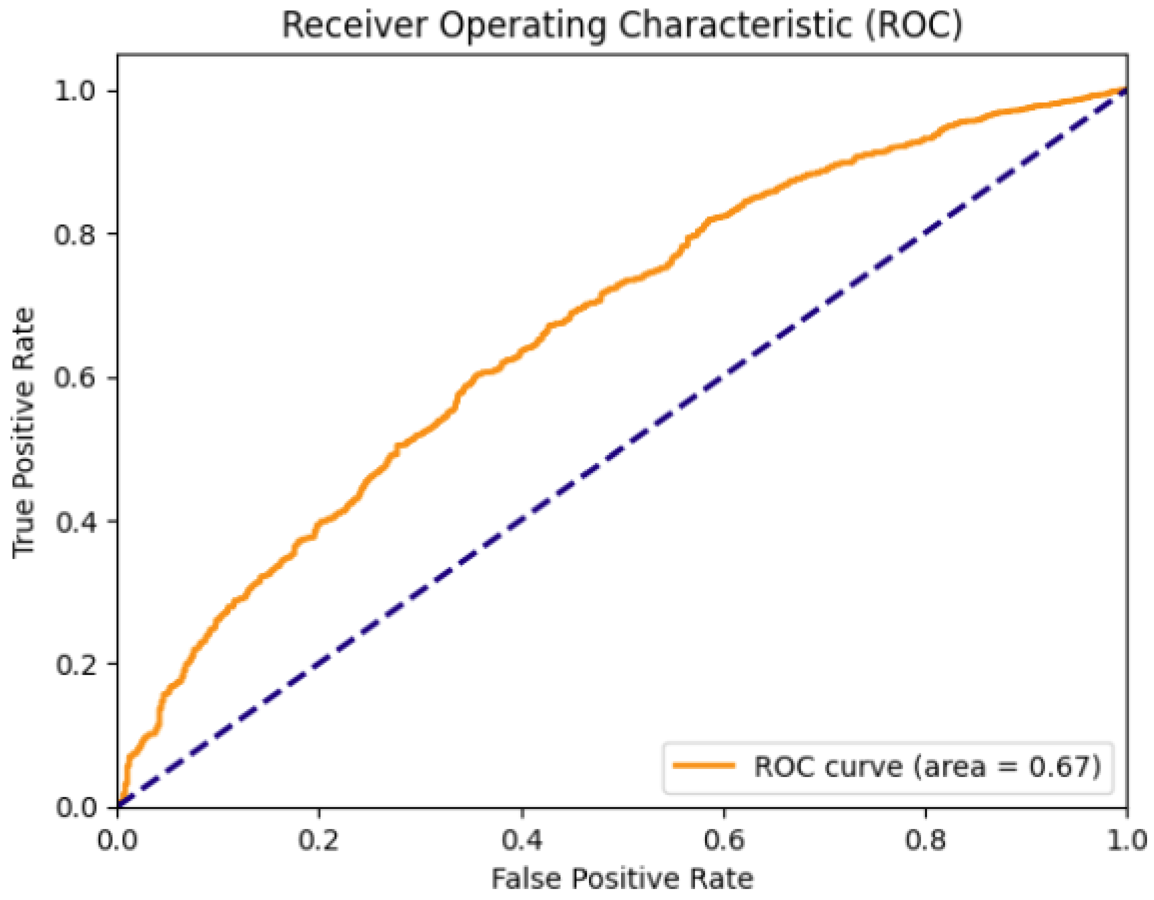
5.2.2.3 Random Forest



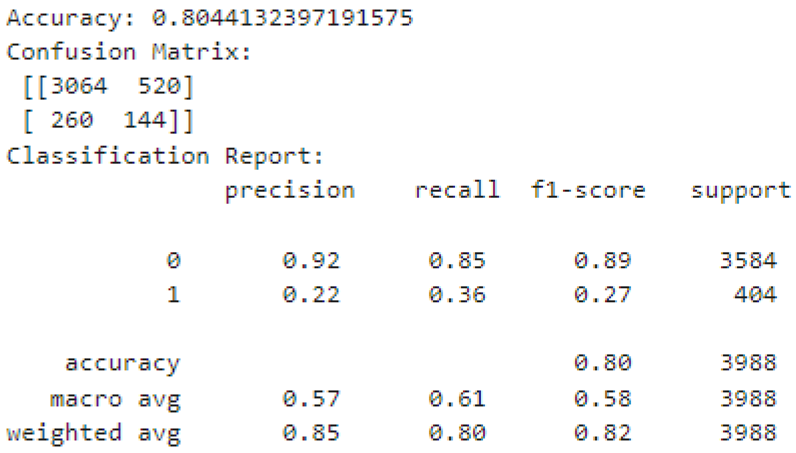
```
Accuracy: 0.7867602808425276
Confusion Matrix:
 [[ 195  306]
 [ 757 3727]]
Classification Report:
               precision    recall  f1-score   support

          -1       0.20      0.39      0.27       501
           1       0.92      0.83      0.88      4484

    accuracy                           0.79      4985
   macro avg       0.56      0.61      0.57      4985
weighted avg       0.85      0.79      0.81      4985
```

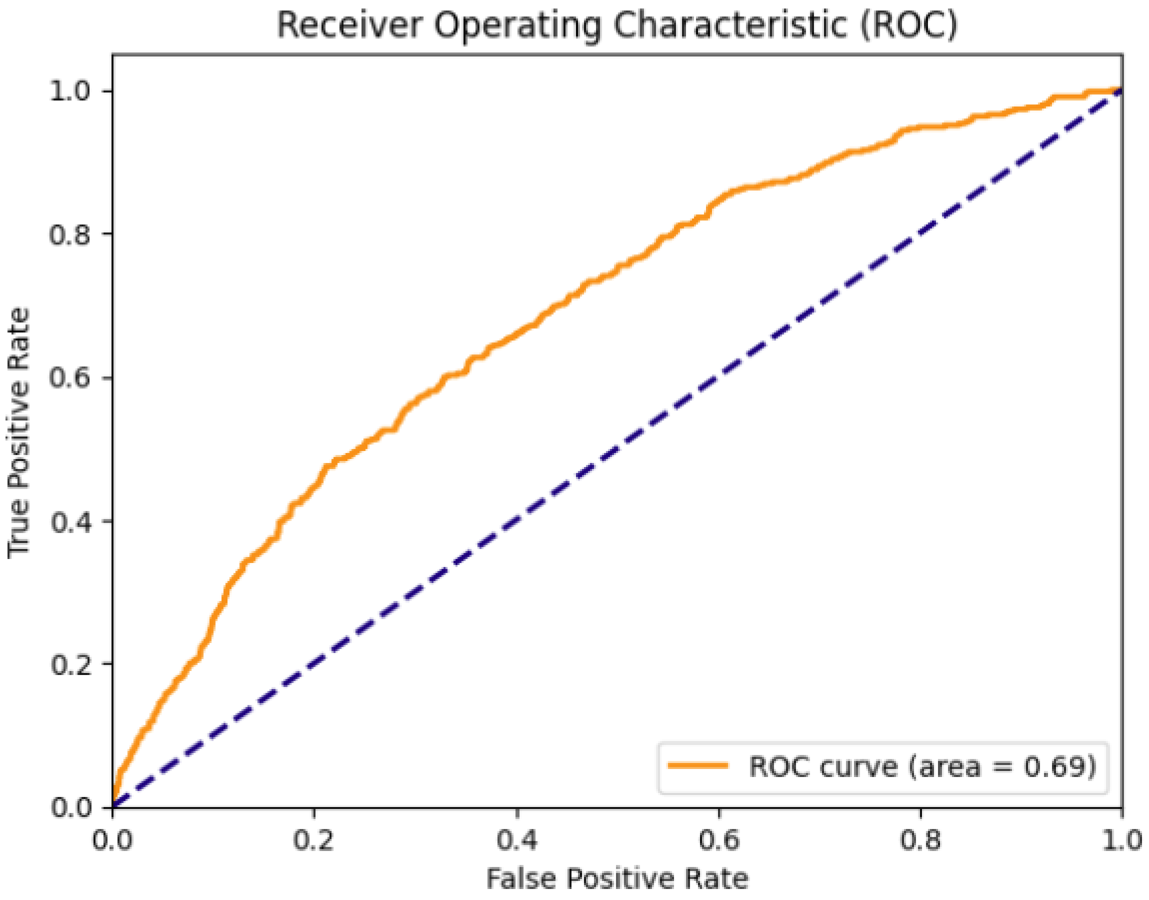


5.2.2.4 Naïve Bayes

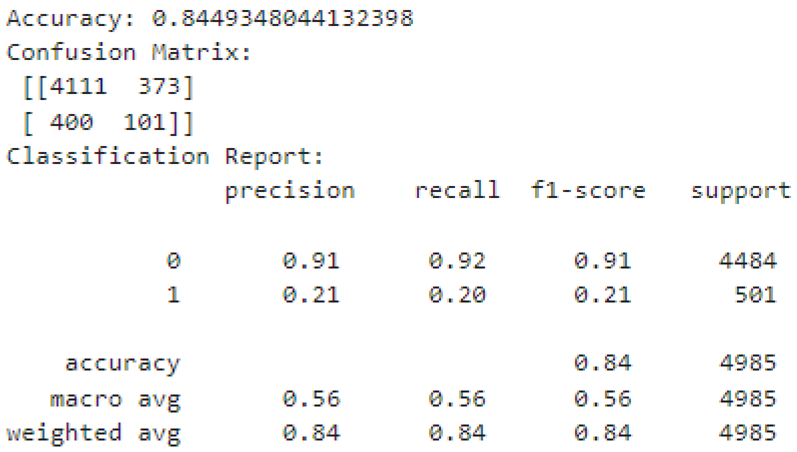

```
Accuracy: 0.8044132397191575
Confusion Matrix:
 [[3064  520]
 [ 260  144]]
Classification Report:
               precision    recall  f1-score   support

           0       0.92      0.85      0.89      3584
           1       0.22      0.36      0.27       404

    accuracy                           0.80      3988
   macro avg       0.57      0.61      0.58      3988
weighted avg       0.85      0.80      0.82      3988
```



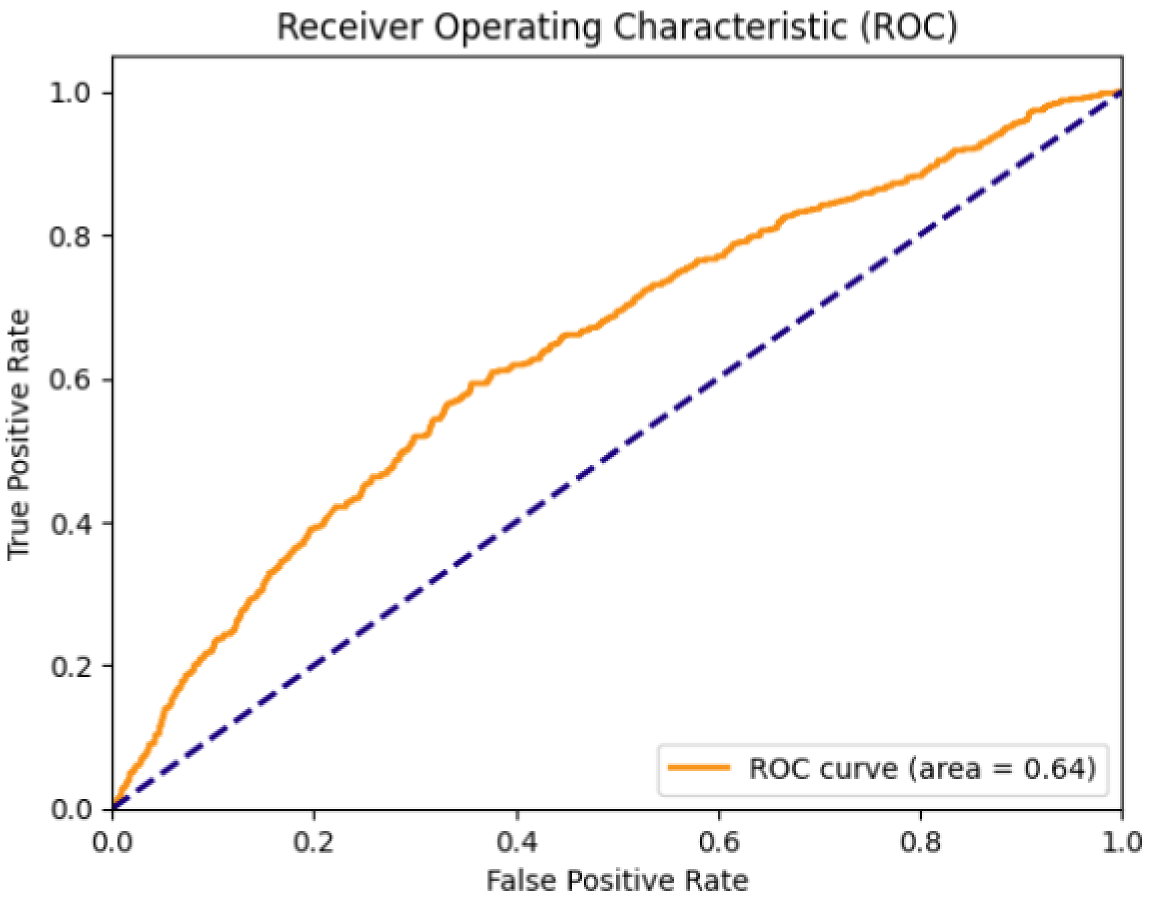
5.2.2.5 XGBoost



```
Accuracy: 0.8449348044132398
Confusion Matrix:
 [[4111  373]
 [ 400  101]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.92      0.91      4484
           1       0.21      0.20      0.21       501

    accuracy                           0.84      4985
   macro avg       0.56      0.56      0.56      4985
weighted avg       0.84      0.84      0.84      4985
```



We have evaluated various ML Learning classifiers with TFIDF embedding technique. XGBoost has been found to outperform all other classifiers. SVM Classifier is found to perform best in specificity.

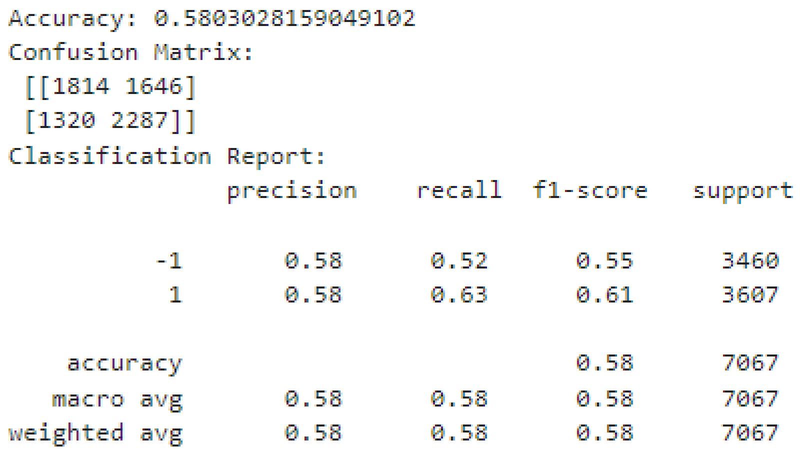
Table 2 TFIDF Embedding Technique comparison with ML Classifiers

TFIDF Embeddings										
Machine Learning Models					**Genuine Review**			**Fake Review**		
	Sensitivity	Specificity	Accuracy%	ROC Area	precision %	recall %	F1 Score %	precision %	recall %	F1 Score %
Logistics Regression Classifier	84.03%	35.52%	79.15%	0.67	92.00%	84.00%	88.00%	20.00%	36.00%	26.00%
Random Forest	83.11%	38.92%	78.67%	0.67	92.00%	83.00%	88.00%	20.00%	39.00%	27.00%
SVM Classifier (Linear)	76.69%	41.11%	73.11%	0.65	92.00%	77.00%	84.00%	16.00%	41.00%	24.00%
Naïve Bayes Classifier	85.46%	35.64%	80.44%	0.69	92.00%	85.00%	89.00%	22.00%	36.00%	27.00%
XGBoost	91.61%	20.15%	84.49%	0.64	91.00%	92.00%	91.00%	21.00%	20.00%	21.00%

5.2.2 Glove Embedding Comparison with Machine Learning Classifiers

Here we will discuss the results of various metrics captured for each ML Models.

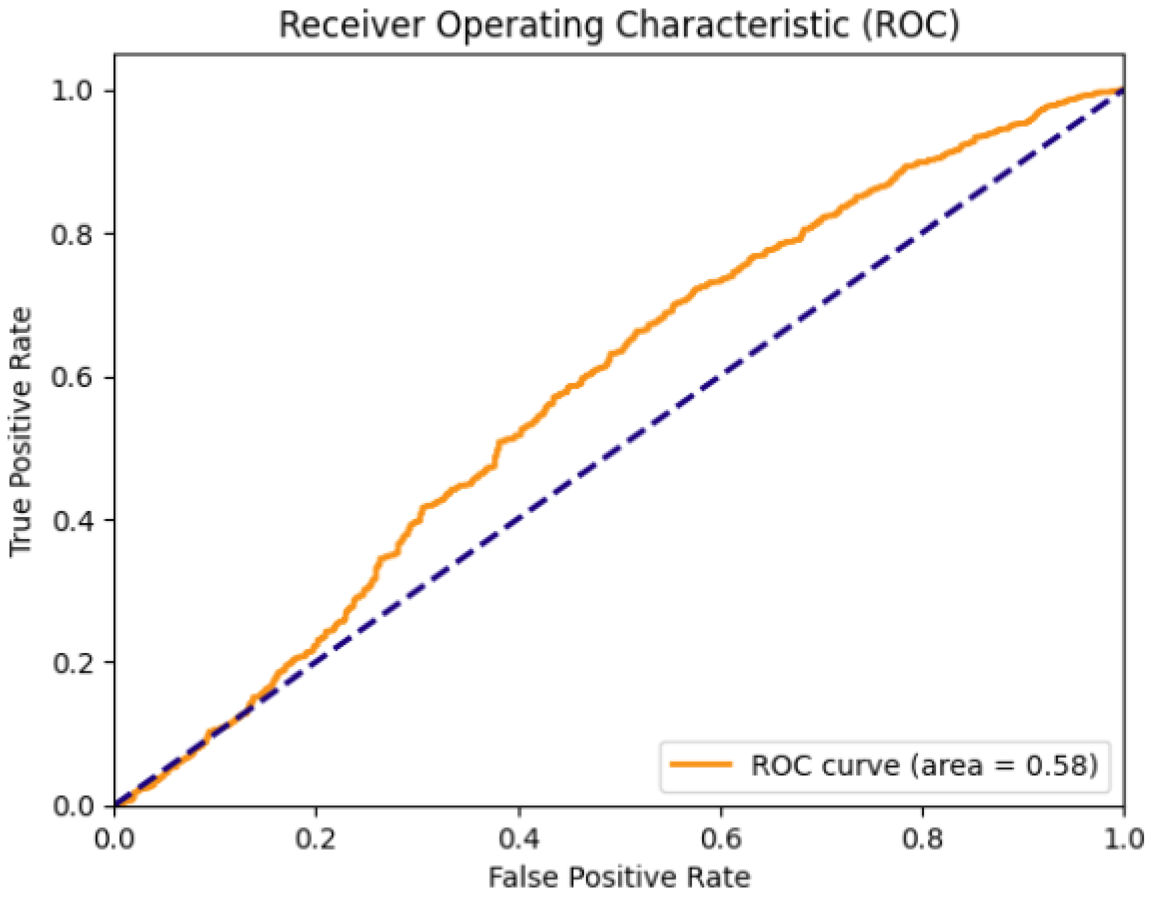
5.2.2.1 Logistic Regression



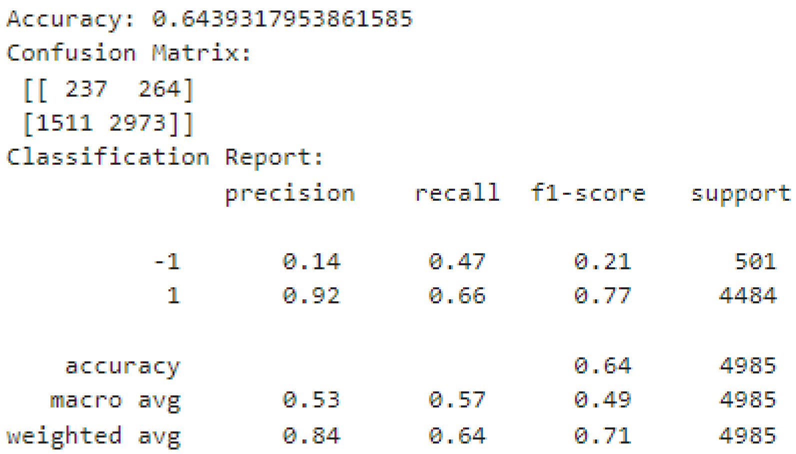
```
Accuracy: 0.5803028159049102
Confusion Matrix:
 [[1814 1646]
 [1320 2287]]
Classification Report:
               precision    recall  f1-score   support

          -1       0.58      0.52      0.55      3460
           1       0.58      0.63      0.61      3607

    accuracy                           0.58      7067
   macro avg       0.58      0.58      0.58      7067
weighted avg       0.58      0.58      0.58      7067
```



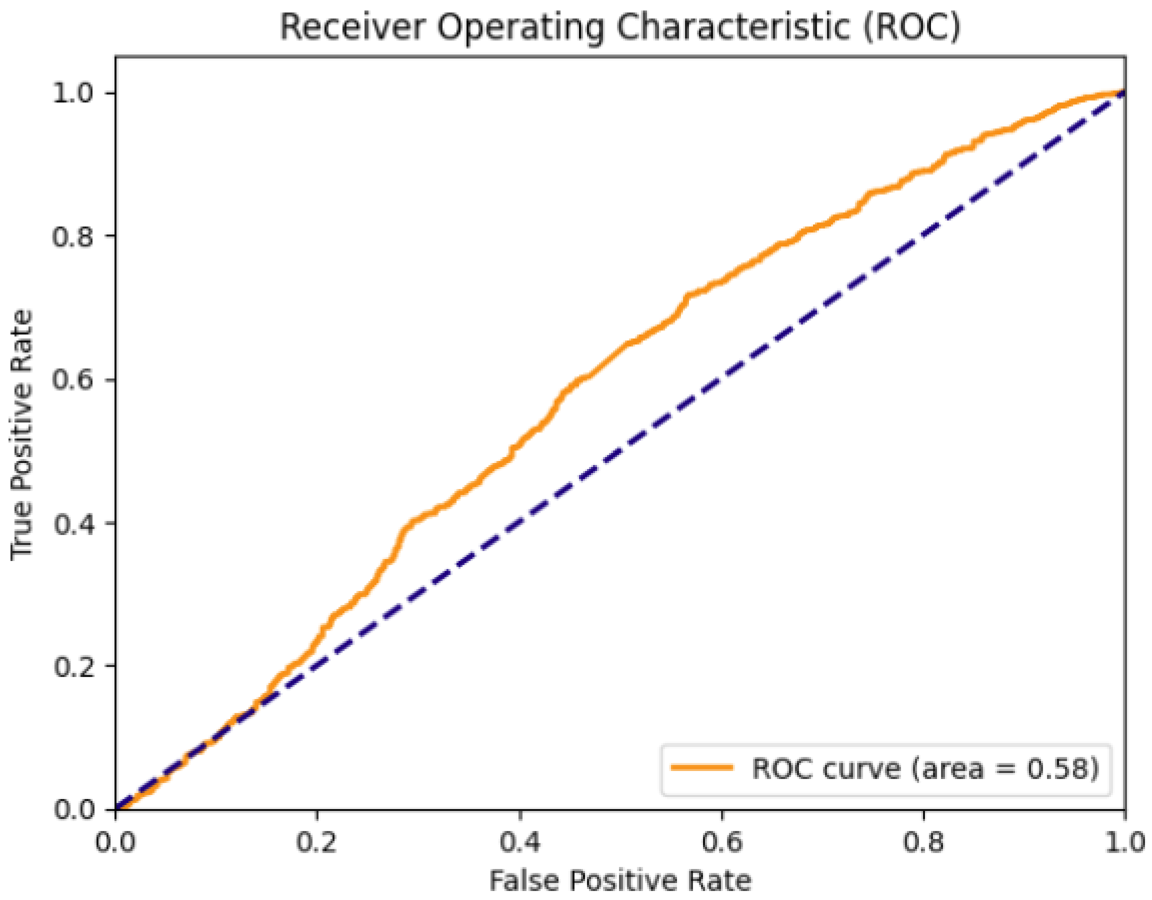
5.2.2.2 Support Vector Machine



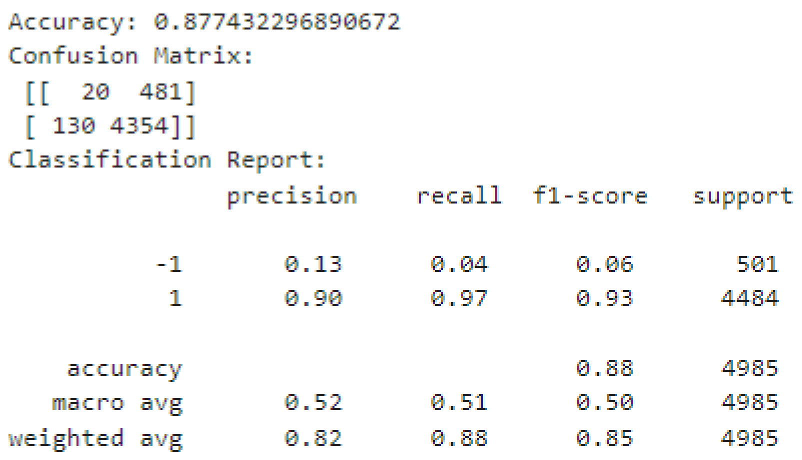
```
Accuracy: 0.6439317953861585
Confusion Matrix:
 [[ 237  264]
 [1511 2973]]
Classification Report:
              precision    recall  f1-score   support

          -1       0.14      0.47      0.21       501
           1       0.92      0.66      0.77      4484

    accuracy                           0.64      4985
   macro avg       0.53      0.57      0.49      4985
weighted avg       0.84      0.64      0.71      4985
```



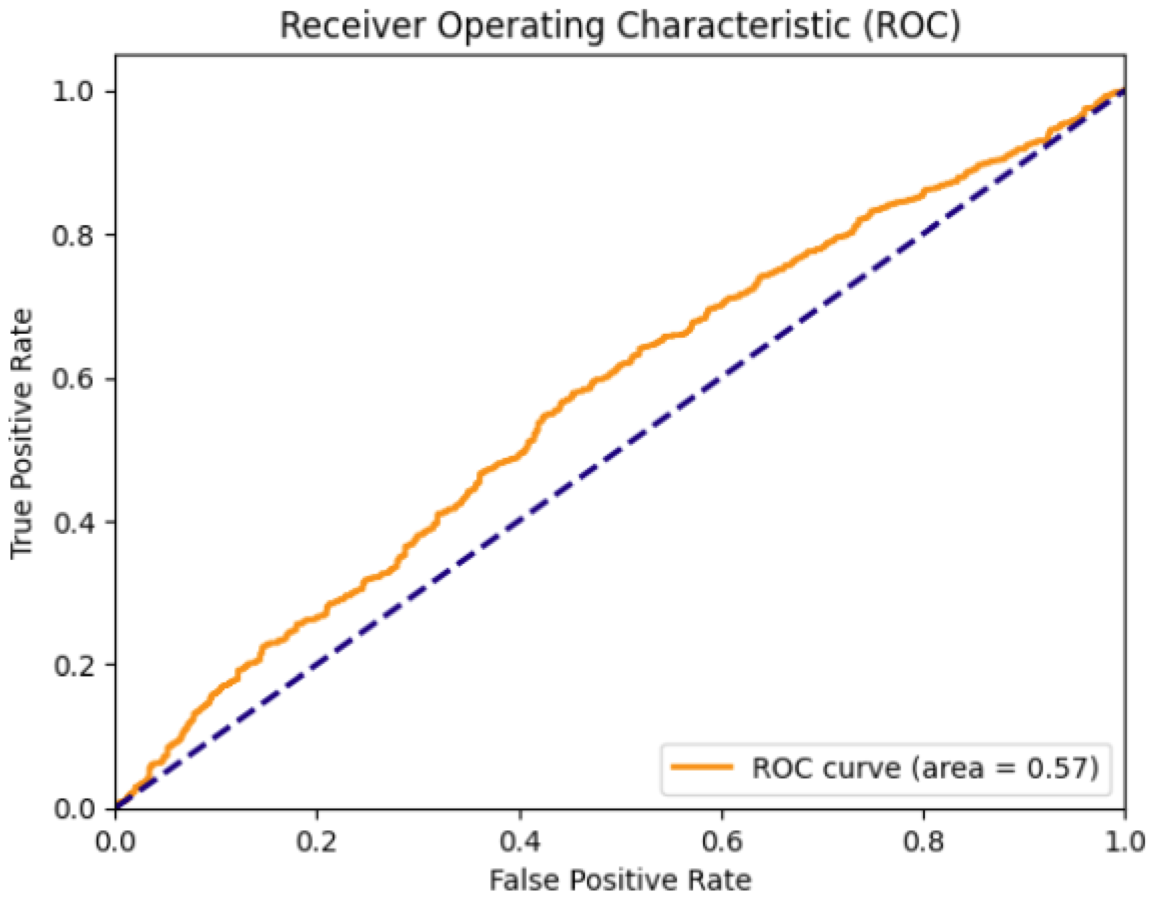
5.2.2.3 Random Forest



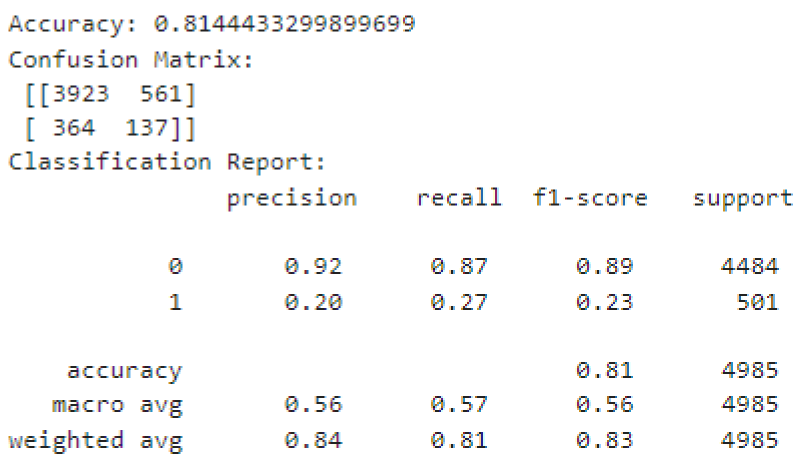
```
Accuracy: 0.877432296890672
Confusion Matrix:
 [[  20  481]
 [ 130 4354]]
Classification Report:
               precision    recall  f1-score   support

          -1       0.13      0.04      0.06       501
           1       0.90      0.97      0.93      4484

    accuracy                           0.88      4985
   macro avg       0.52      0.51      0.50      4985
weighted avg       0.82      0.88      0.85      4985
```



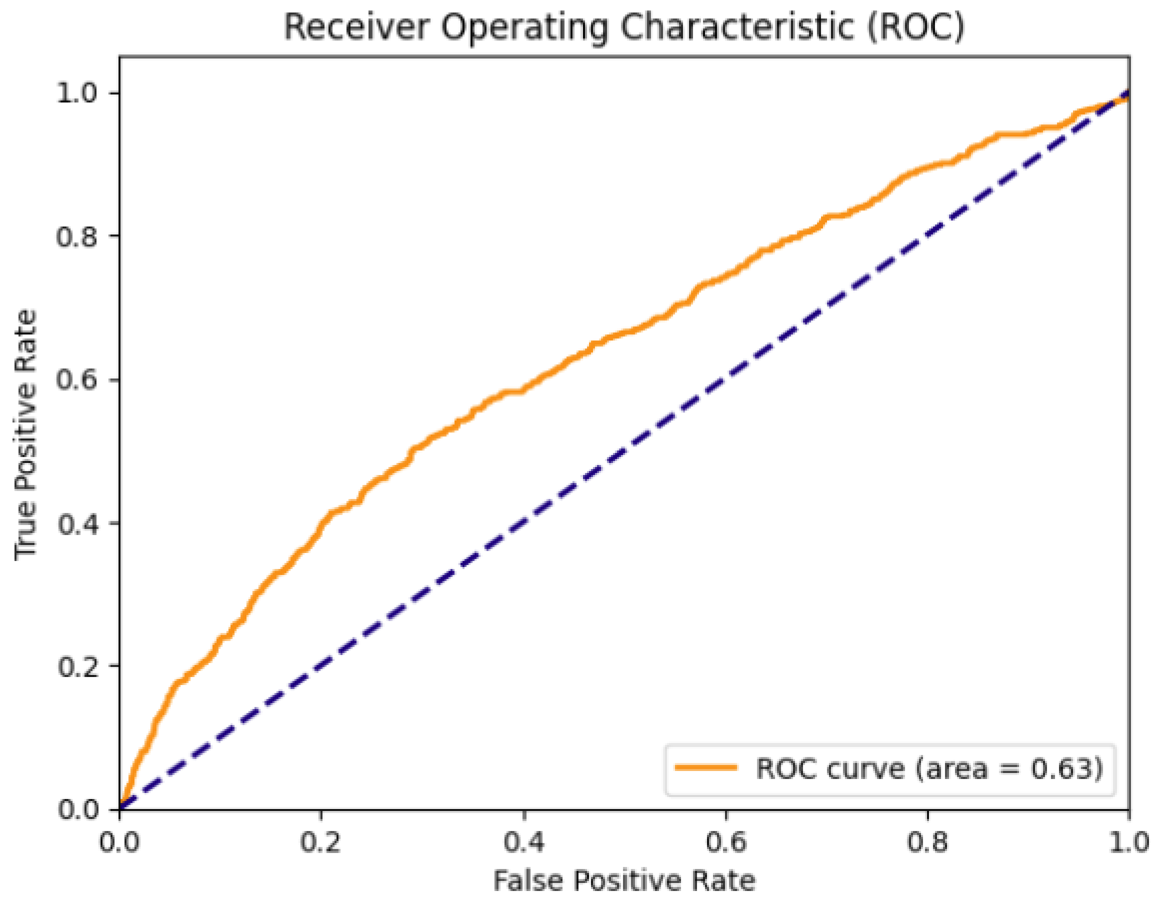
5.2.2.4 Naïve Bayes



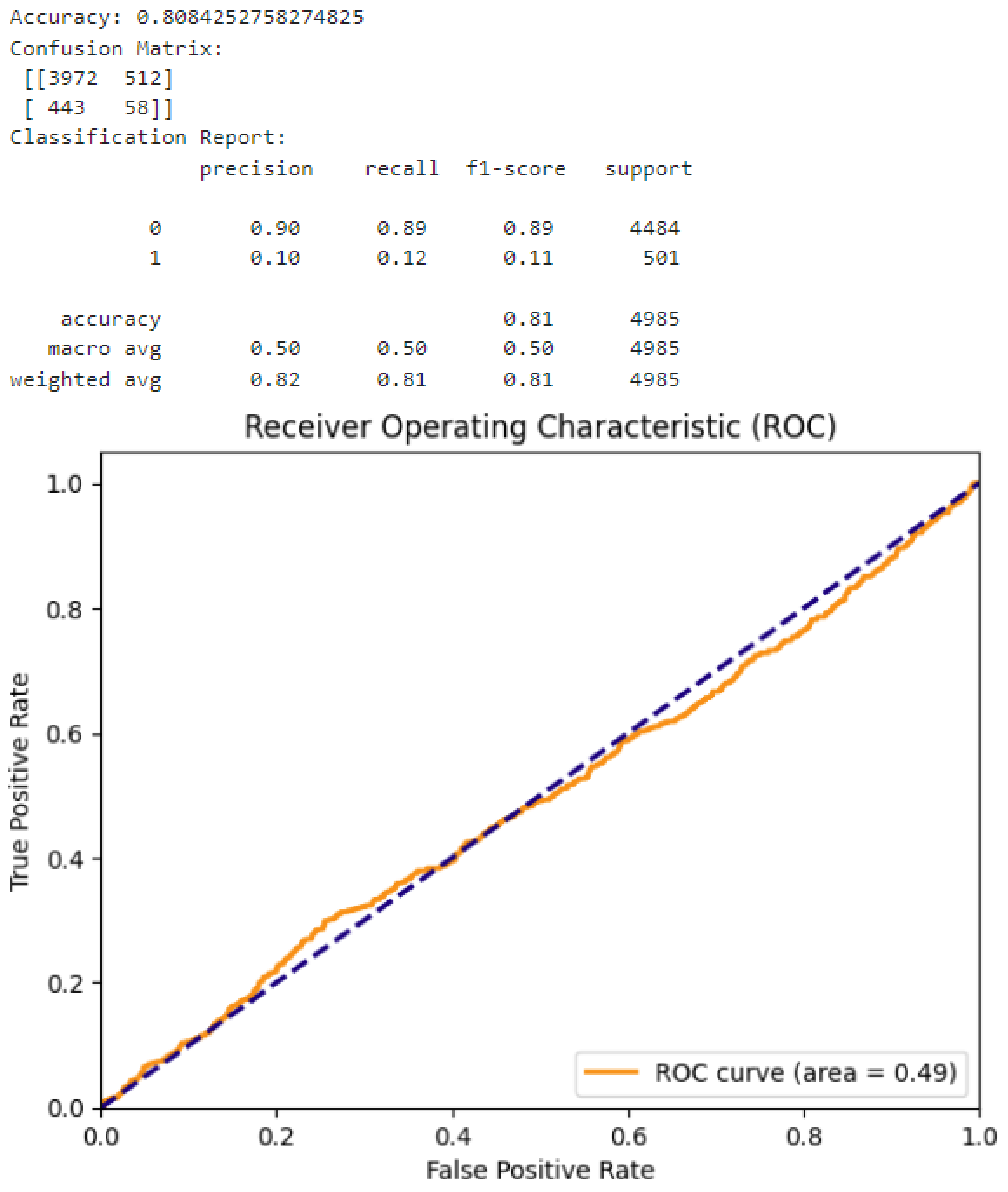
```
Accuracy: 0.8144433299899699
Confusion Matrix:
 [[3923  561]
 [ 364  137]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.87      0.89      4484
           1       0.20      0.27      0.23       501

    accuracy                           0.81      4985
   macro avg       0.56      0.57      0.56      4985
weighted avg       0.84      0.81      0.83      4985
```



5.2.2.5 XGBoost



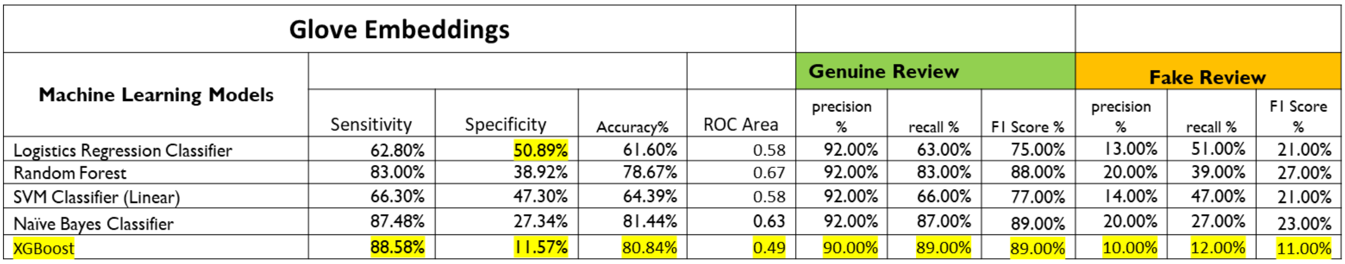
```
Accuracy: 0.8084252758274825
Confusion Matrix:
 [[3972  512]
 [ 443   58]]
Classification Report:
               precision    recall  f1-score   support

           0       0.90      0.89      0.89      4484
           1       0.10      0.12      0.11       501

    accuracy                           0.81      4985
   macro avg       0.50      0.50      0.50      4985
weighted avg       0.82      0.81      0.81      4985
```

We have evaluated various ML Learning classifiers with Glove embedding technique. XGBoost has been found to outperform all other classifiers. Logistic Regression Classifier is found to perform best in specificity.

Table 3 Glove Embedding Technique comparison with ML Classifiers

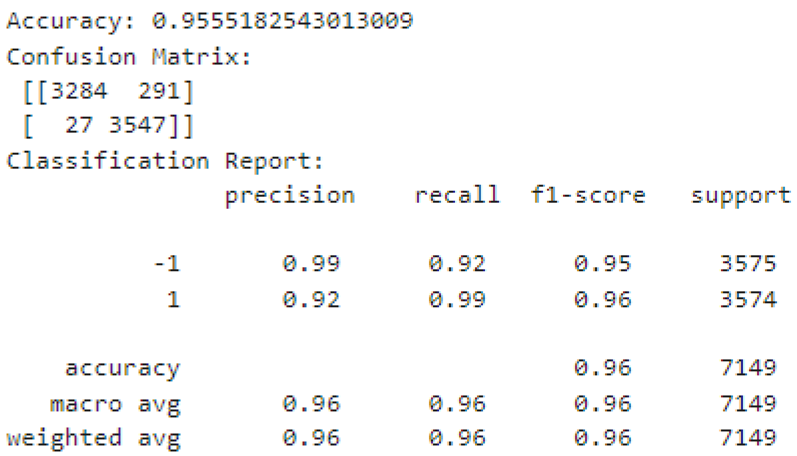


Glove Embeddings										
Machine Learning Models					Genuine Review			Fake Review		
	Sensitivity	Specificity	Accuracy%	ROC Area	precision %	recall %	F1 Score %	precision %	recall %	F1 Score %
Logistics Regression Classifier	62.80%	50.89%	61.60%	0.58	92.00%	63.00%	75.00%	13.00%	51.00%	21.00%
Random Forest	83.00%	38.92%	78.67%	0.67	92.00%	83.00%	88.00%	20.00%	39.00%	27.00%
SVM Classifier (Linear)	66.30%	47.30%	64.39%	0.58	92.00%	66.00%	77.00%	14.00%	47.00%	21.00%
Naïve Bayes Classifier	87.48%	27.34%	81.44%	0.63	92.00%	87.00%	89.00%	20.00%	27.00%	23.00%
XGBoost	88.58%	11.57%	80.84%	0.49	90.00%	89.00%	89.00%	10.00%	12.00%	11.00%

5.2.3 BERT based embedding Comparison with ML and DL Classifiers

Here we will discuss the results of various metrics captured for each ML and DL Models with BERT, RoBERTa, DistilBERT, XLNet embedding techniques.

5.2.3.1 BERT with SVM Classifier

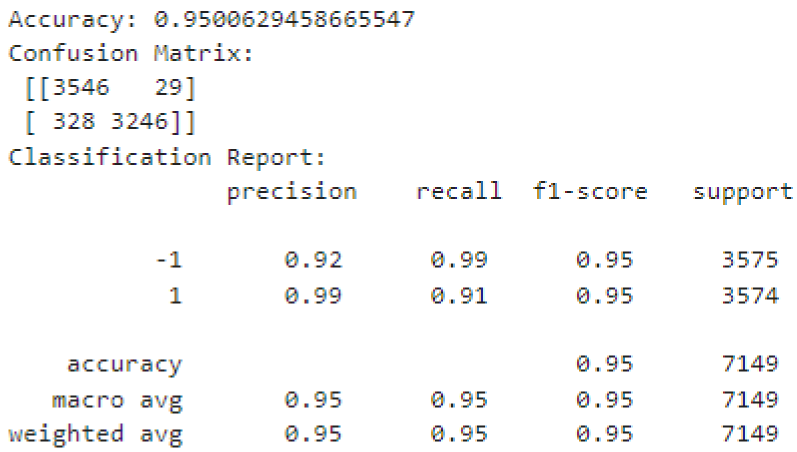


```
Accuracy: 0.9555182543013009
Confusion Matrix:
 [[3284  291]
 [  27 3547]]
Classification Report:
              precision    recall  f1-score   support

          -1       0.99      0.92      0.95      3575
           1       0.92      0.99      0.96      3574

    accuracy                           0.96      7149
   macro avg       0.96      0.96      0.96      7149
weighted avg       0.96      0.96      0.96      7149
```

5.2.3.2 BERT with MLP Classifier

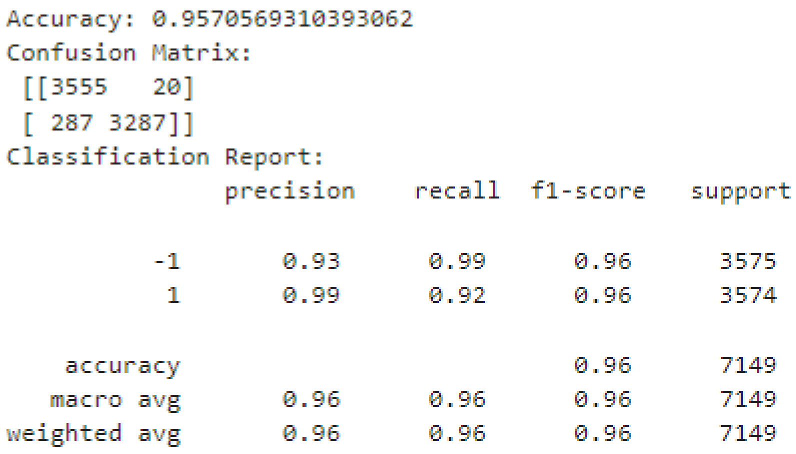


```
Accuracy: 0.9500629458665547
Confusion Matrix:
 [[3546   29]
 [ 328 3246]]
Classification Report:
              precision    recall  f1-score   support

          -1       0.92      0.99      0.95      3575
           1       0.99      0.91      0.95      3574

    accuracy                           0.95      7149
   macro avg       0.95      0.95      0.95      7149
weighted avg       0.95      0.95      0.95      7149
```

5.2.3.3 RoBERTa with MLP Classifier

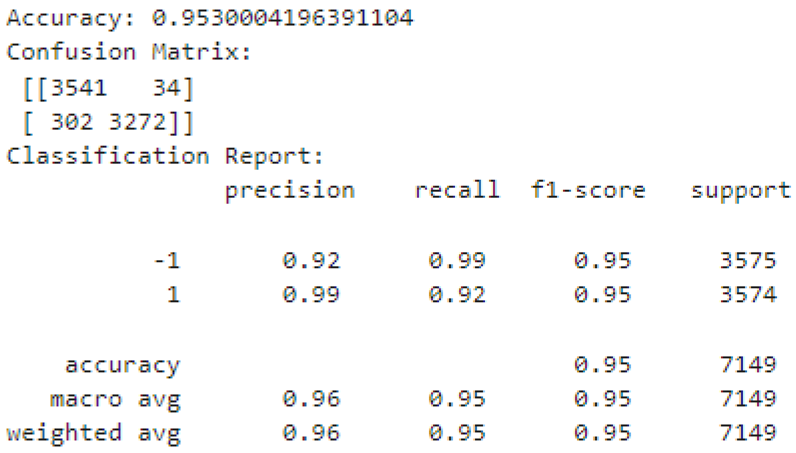


```
Accuracy: 0.9570569310393062
Confusion Matrix:
 [[3555   20]
 [ 287 3287]]
Classification Report:
              precision    recall  f1-score   support

          -1       0.93      0.99      0.96      3575
           1       0.99      0.92      0.96      3574

    accuracy                           0.96      7149
   macro avg       0.96      0.96      0.96      7149
weighted avg       0.96      0.96      0.96      7149
```

5.2.3.4 DistilBERT with MLP Classifier

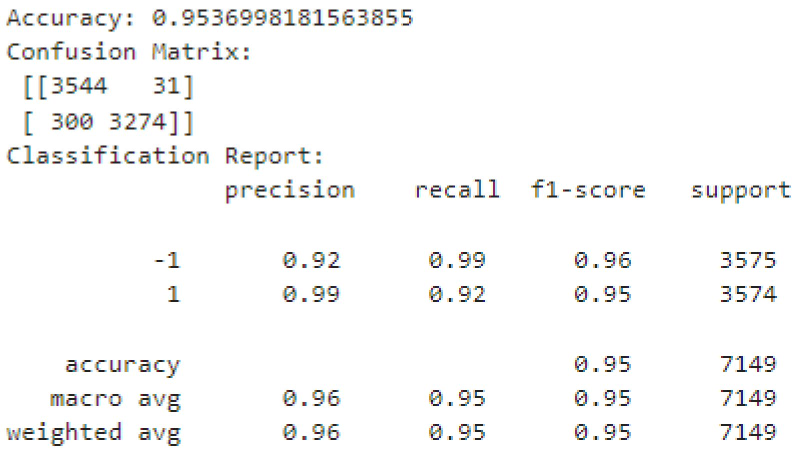


```
Accuracy: 0.9530004196391104
Confusion Matrix:
 [[3541   34]
 [ 302 3272]]
Classification Report:
              precision    recall  f1-score   support

          -1       0.92      0.99      0.95      3575
           1       0.99      0.92      0.95      3574

    accuracy                           0.95      7149
   macro avg       0.96      0.95      0.95      7149
weighted avg       0.96      0.95      0.95      7149
```

5.2.3.5 XLNet with MLP CLassifier



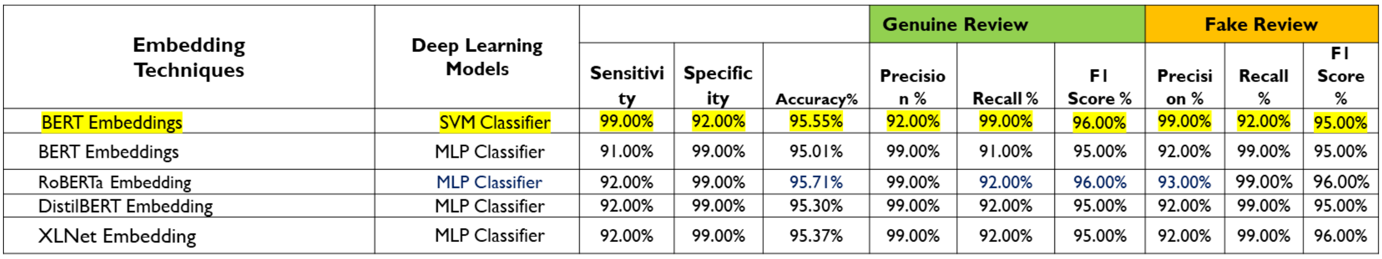
```
Accuracy: 0.9536998181563855
Confusion Matrix:
 [[3544   31]
 [ 300 3274]]
Classification Report:
              precision    recall  f1-score   support

          -1       0.92      0.99      0.96      3575
           1       0.99      0.92      0.95      3574

    accuracy                           0.95      7149
   macro avg       0.96      0.95      0.95      7149
weighted avg       0.96      0.95      0.95      7149
```

We have evaluated various ML and DL Learning classifiers with BERT, RoBERTa, DistilBERT, XLNet embedding techniques. **SVM Classifier** has been found to outperform all other classifiers.

Table 4 BERT, RoBERTa, DistilBERT and XLNet Embedding comparison across ML/DL Models

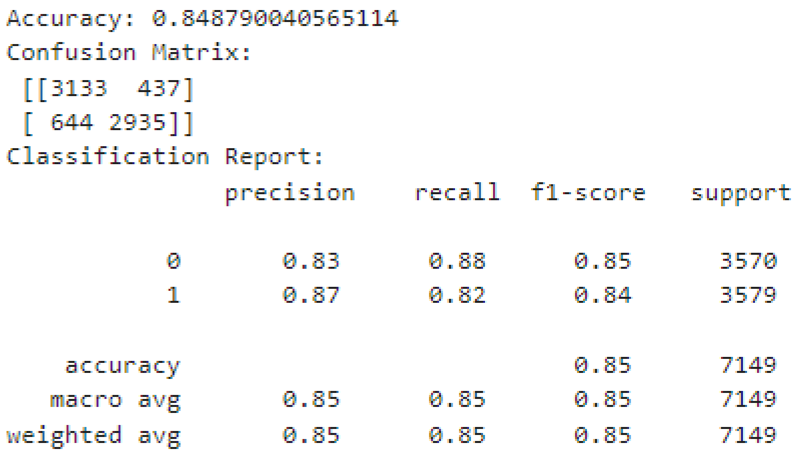


Embedding Techniques	Deep Learning Models				Genuine Review			Fake Review		
		Sensitivity	Specificity	Accuracy%	Precision %	Recall %	F1 Score %	Precision %	Recall %	F1 Score %
BERT Embeddings	SVM Classifier	99.00%	92.00%	95.55%	92.00%	99.00%	96.00%	99.00%	92.00%	95.00%
BERT Embeddings	MLP Classifier	91.00%	99.00%	95.01%	99.00%	91.00%	95.00%	92.00%	99.00%	95.00%
RoBERTa Embedding	MLP Classifier	92.00%	99.00%	95.71%	99.00%	92.00%	96.00%	93.00%	99.00%	96.00%
DistilBERT Embedding	MLP Classifier	92.00%	99.00%	95.30%	99.00%	92.00%	95.00%	92.00%	99.00%	95.00%
XLNet Embedding	MLP Classifier	92.00%	99.00%	95.37%	99.00%	92.00%	95.00%	92.00%	99.00%	96.00%

5.2.4 Comparison of Deep Learning Classifiers

Here we will discuss the results of various metrics captured for each DL Models

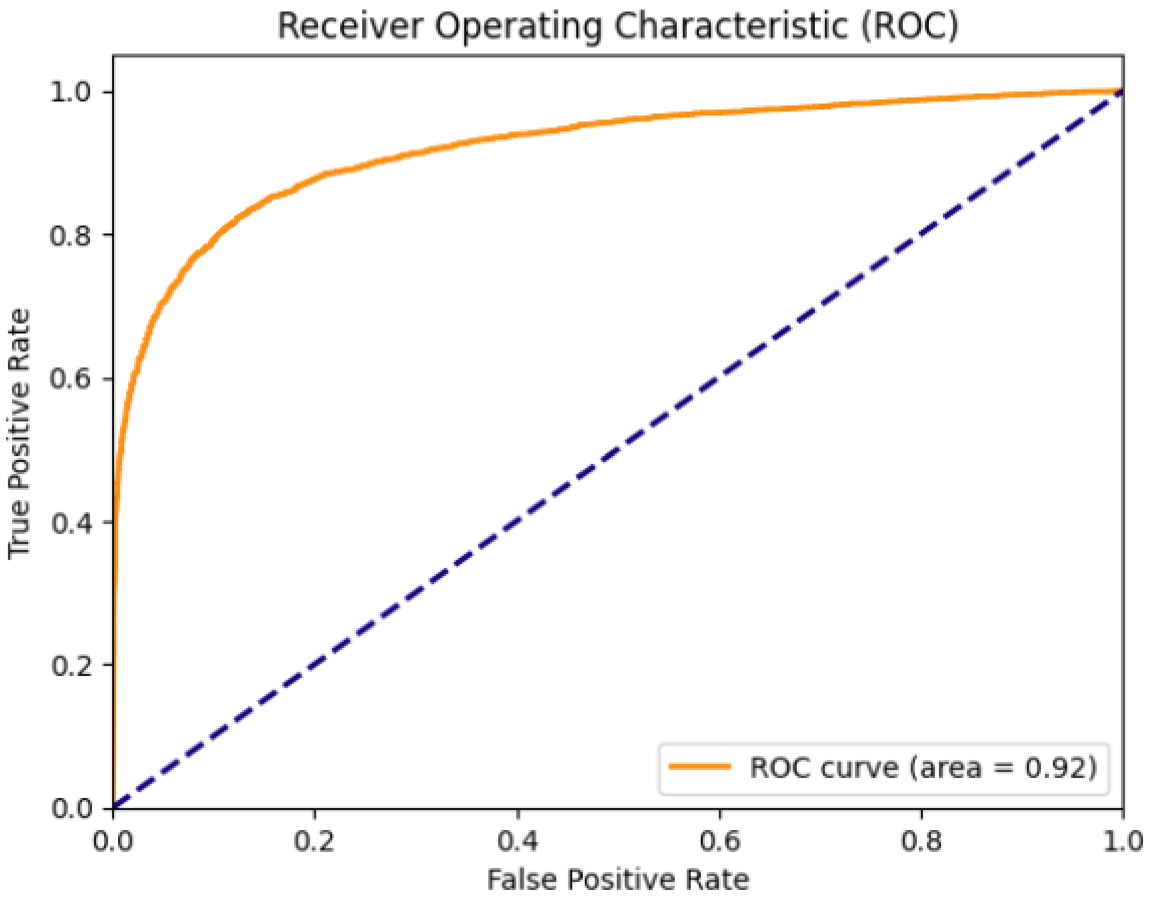
5.2.4.1 CNN



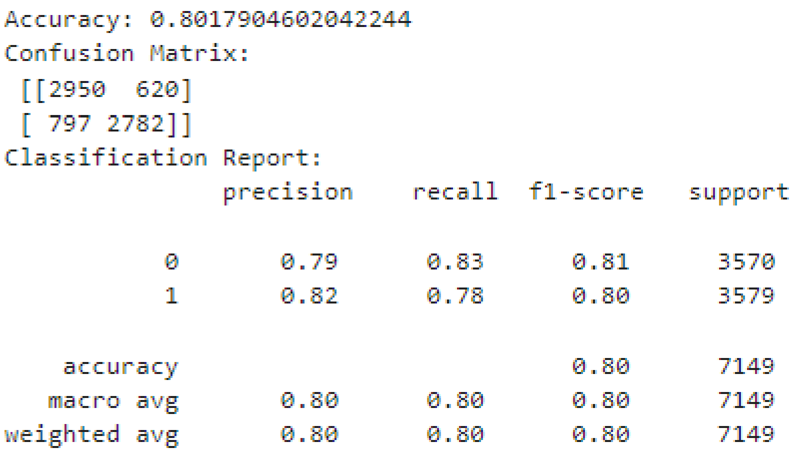
```
Accuracy: 0.848790040565114
Confusion Matrix:
 [[3133  437]
 [ 644 2935]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.88      0.85      3570
           1       0.87      0.82      0.84      3579

    accuracy                           0.85      7149
   macro avg       0.85      0.85      0.85      7149
weighted avg       0.85      0.85      0.85      7149
```



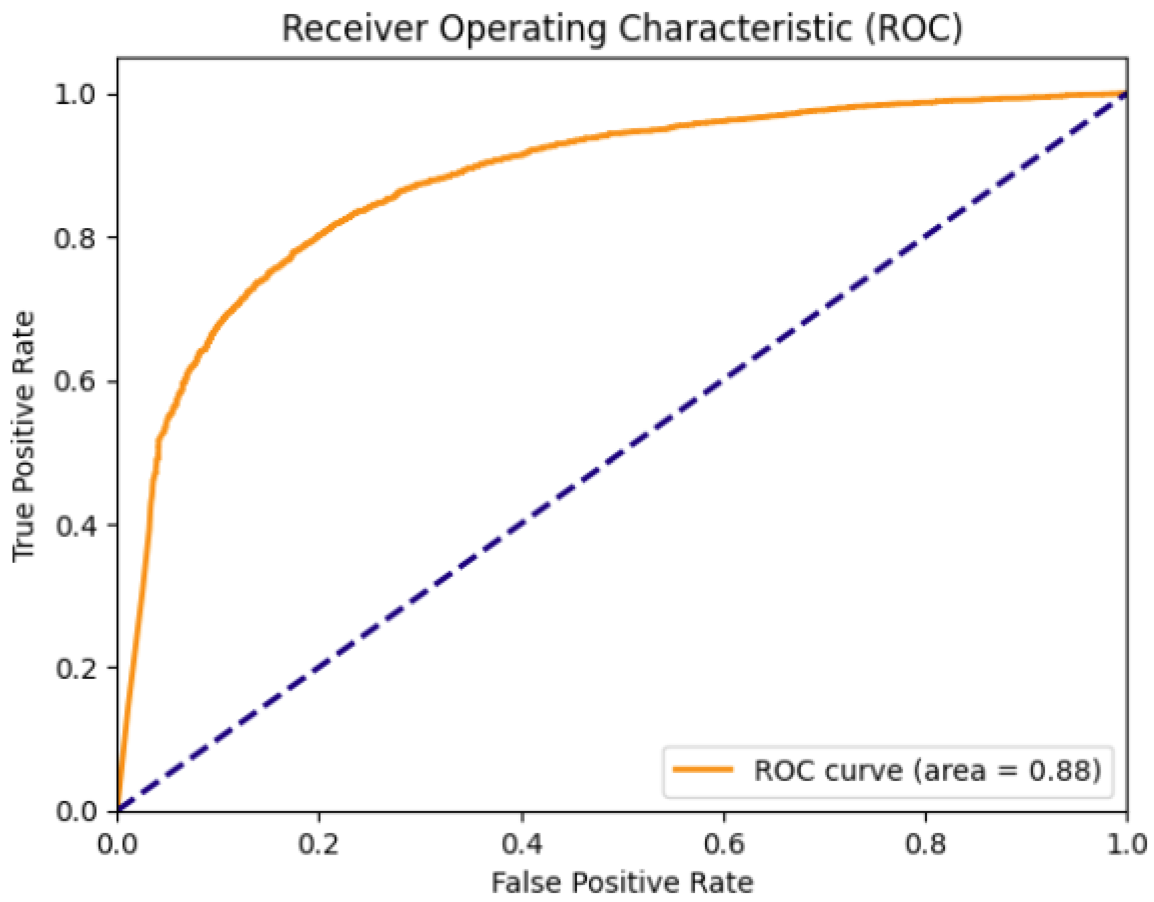
5.2.4.2 CNN-LSTM



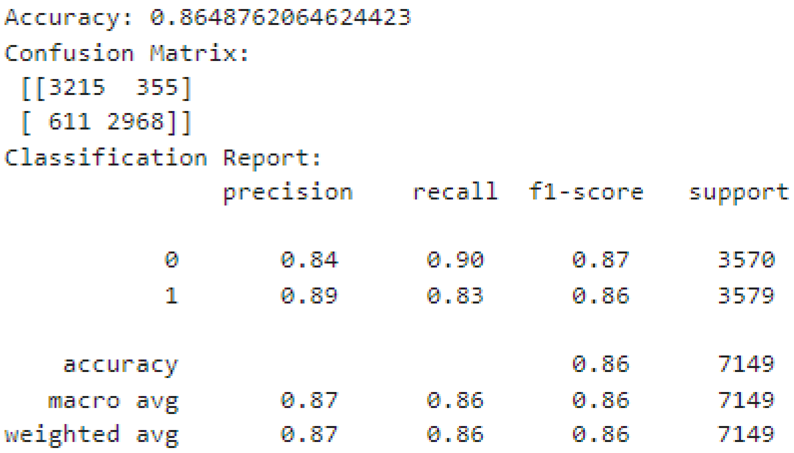
```
Accuracy: 0.8017904602042244
Confusion Matrix:
 [[2950  620]
 [ 797 2782]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.83      0.81      3570
           1       0.82      0.78      0.80      3579

    accuracy                           0.80      7149
   macro avg       0.80      0.80      0.80      7149
weighted avg       0.80      0.80      0.80      7149
```



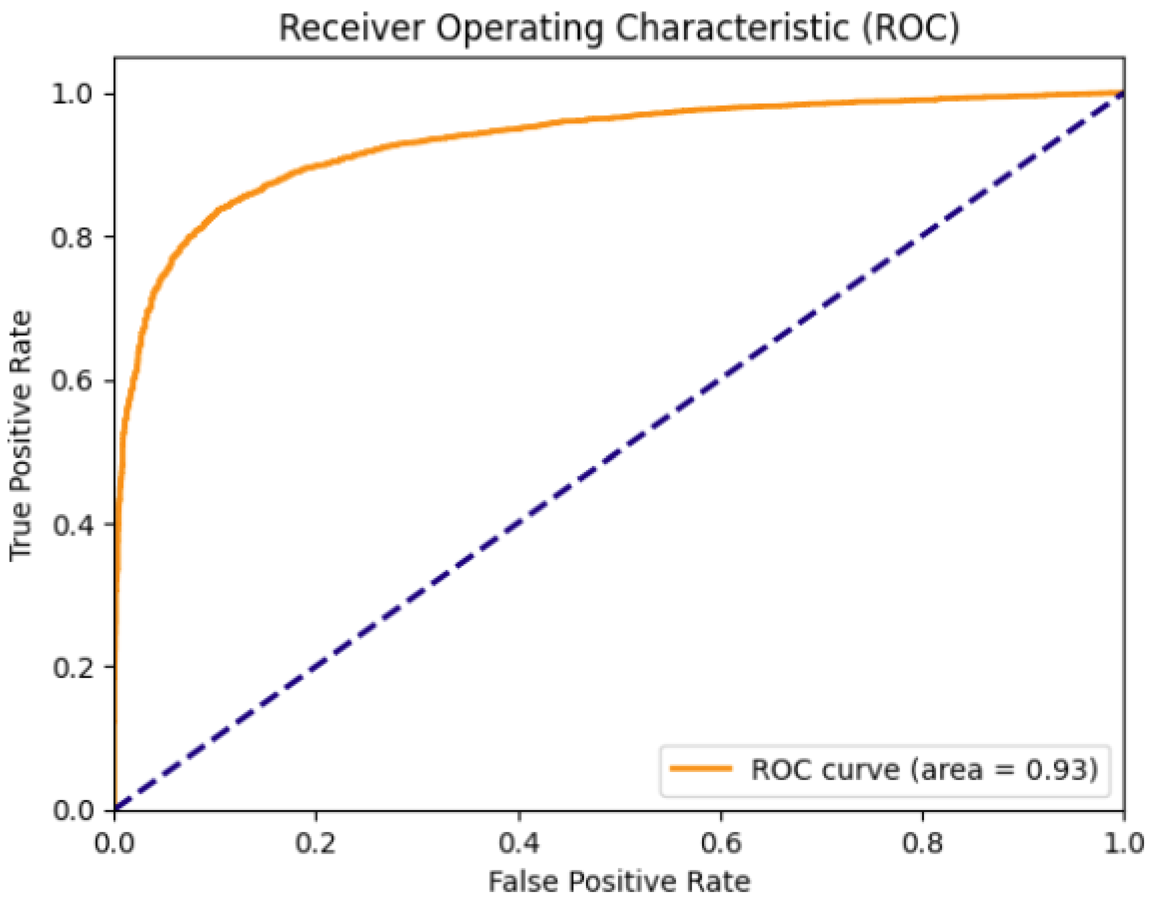
5.2.4.3 RNN-LSTM



```
Accuracy: 0.8648762064624423
Confusion Matrix:
 [[3215  355]
 [ 611 2968]]
Classification Report:
               precision    recall  f1-score   support

           0       0.84      0.90      0.87      3570
           1       0.89      0.83      0.86      3579

    accuracy                           0.86      7149
   macro avg       0.87      0.86      0.86      7149
weighted avg       0.87      0.86      0.86      7149
```



We have evaluated fake review detection with Deep Learning classifiers such as CNN, CNN-LSTM, RNN-LSTM. RNN-LSTM Classifier is found to out perform all others.

Table 5 Deep Learning Classifiers Comparison

Deep Learning Models				Genuine Review			Fake Review		
	Sensitivity	Specificity	Accuracy %	precision %	recall %	F1 Score %	precision %	recall %	F1 Score %
CNN Classifier	88.00%	92.00%	84.88%	83.00%	88.00%	85.00%	99.00%	92.00%	95.00%
CNN + LSTM Classifier	83.00%	78.00%	80.18%	79.00%	83.00%	81.00%	82.00%	78.00%	80.00%
RNN LSTM Classifier	90.00%	83.00%	86.49%	84.00%	90.00%	87.00%	89.00%	83.00%	86.00%

5.3 Model Prediction with Live Data:

Once the Models are trained and tested, model is saved and used to predict Labels on new YELP data set. For this purpose a separate sample data set is prepared from Yelp Business dataset. Results are found as expected.

5.3.1 Testing of Deep Learning Models on Live Yelp data set :

Once the Models are tested, it is saved and tested on YELP Live data set. For this purpose a separate sample data set is prepared from Yelp. The results are validated and found as expected.

5.4 Summary

It is observed **BERT** based Embedding techniques are clearly outperforming **TFIDF and Glove** Embedding techniques.

BERT and its variants **RoBERTa and DistilBERT, XLNET** with contextual word embeddings are able to outperform traditional embeddings such as TFIDF and Glove embeddings which captures word importance and semantic importance.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

- Various Embedding techniques **TFIDF, Glove, BERT, RoBERTa, DistilBERT and XLNET** along with Machine Learning and Deep Learning models were tested and evaluated for Fake Review Detection. It is found that Deep Learning techniques with BERT and BERT based embeddings techniques are giving much better results.
- These techniques are outperforming traditional embedding techniques such as TFIDF, Word2Vec and Glove.
- Processing Time for model training using Deep Learning techniques-based model is significantly large than processing time for model training with Machine Learning models. For eg. Training the machine learning model takes around 15-20 minutes. Whereas training deep learning based model takes around 3-4 hours. This is due to additional complexity involved in neural networks and transformer based pre-trained model.
- Due to resource and processor limitation, selected experimentation were done on different embedding techniques such as **TFIDF, Glove, BERT, RoBERTa, DistilBERT and XLNET**.
- Future work can focus on fine tuning these Embedding techniques in Deep Neural Networks improving the fake review detection methods using Unsupervised and Semi supervised models.

Based on findings in the thesis we are successful to come up with an approach to identify fake reviews and genuine reviews to meet the Aim of this Thesis.

We are able to analyse various embedding techniques along with Advanced ML and DL models for effective fake review detection.

We are able to propose a suitable model architecture for the fake review detection.

We are able successfully to assess the effectiveness of different embedding techniques such as TFIDF, Glove, BERT, RoBERTa, DistilBERT, XLNET along with ML and DL Models.

We are able to successfully evaluate the overall performance of the proposed framework using appropriate metrics and validation techniques to assess its efficiency and accuracy in fake review detection.

REFERENCES

Aghakhani, H., MacHiry, A., Nilizadeh, S., Kruegel, C. and Vigna, G., (2018) Detecting deceptive reviews using generative adversarial networks. In: *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*. Institute of Electrical and Electronics Engineers Inc., pp.89–95.

Alghamdi, J., Lin, Y. and Luo, S., (2022) A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection. *Information (Switzerland)*, 1312.

Al-Zoubi, A.M., Mora, A.M. and Faris, H., (2023) A Multilingual Spam Reviews Detection Based on Pre-Trained Word Embedding and Weighted Swarm Support Vector Machines. *IEEE Access*, 11, pp.72250–72271.

Asaad, W.H., Allami, R. and Ali, Y.H., (2023) Fake Review Detection Using Machine Learning. *Revue d'Intelligence Artificielle*, 375.

Berry, S. and Howard, W., (2024) *Fake Google restaurant reviews and the implications for consumers and restaurants*.

Deshai, N. and Bhaskara Rao, B., (2023) Unmasking deception: a CNN and adaptive PSO approach to detecting fake online reviews. *Soft Computing*, 2716, pp.11357–11378.

Deshai, N. and Rao, B.B., (2023) Deep Learning Hybrid Approaches to Detect Fake Reviews and Ratings. *Journal of Scientific and Industrial Research*, 821, pp.120–127.

Devlin, J., Chang, M.-W., Lee, K., Google, K.T. and Language, A.I., (n.d.) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] Available at: https://github.com/tensorflow/tensor2tensor.

Duma, R.A., Niu, Z., Nyamawe, A.S., Tchaye-Kondi, J. and Yusuf, A.A., (2023) A Deep Hybrid Model for fake review detection by jointly leveraging review text, overall ratings, and aspect ratings. *Soft Computing*, 2710, pp.6281–6296.

Elmogy, A.M., Tariq, U., Ibrahim, A. and Mohammed, A., (n.d.) *Fake Reviews Detection using Supervised Machine Learning*. [online] *IJACSA) International Journal of Advanced Computer Science and Applications*, Available at: www.ijacsa.thesai.org.

Fang, Y., Wang, H., Zhao, L., Yu, F. and Wang, C., (2020) Dynamic knowledge graph based fake-review detection. *Applied Intelligence*, 5012, pp.4281–4295.

Ganesh, D., Rao, K.J., Kumar, M.S., Vinitha, M., Anitha, M., Likith, S.S. and Taralitha, R., (2023) Implementation of Novel Machine Learning Methods for Analysis and Detection of Fake Reviews in Social Media. In: *2nd International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2023 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp.243–250.

Gupta, P., (2021) *Leveraging Transfer learning techniques-BERT, RoBERTa, ALBERT and DistilBERT for Fake Review Detection MSc Research Project Data Analytics*.

Hajek, P., Barushka, A. and Munk, M., (2020) Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Computing and Applications*, 3223, pp.17259–17274.

Jain, P.K., Pamula, R. and Ansari, S., (2021) A Supervised Machine Learning Approach for the Credibility Assessment of User-Generated Content. *Wireless Personal Communications*, 1184, pp.2469–2485.

Jindal, N. and Liu, B., (2008) *Opinion Spam and Analysis*. [online] Available at: http://money.cnn.com/2006.

Li Qian and Wu, Q. and Z.C. and Z.J. and Z.W., (2019) Unsupervised User Behavior Representation for Fraud Review Detection with Cold-Start Problem. In: Z.-H. and G.Z. and Z.M.-L. and H.S.-J. Yang Qiang and Zhou, ed., *Advances in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing, pp.222–236.

Manaskasemsak, B., Tantisuwankul, J. and Rungsawang, A., (2023) Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network. *Neural Computing and Applications*, 352, pp.1169–1182.

Mewada, A., Dewang, R.K., Goldar, P. and Maurya, S.K., (2023) SentiBERT: A Novel Approach for Fake Review Detection Incorporating Sentiment Features with Contextual Features. In: *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp.230–235.

Mir, A.Q., Khan, F.Y. and Chishti, M.A., (n.d.) *ONLINE FAKE REVIEW DETECTION USING SUPERVISED MACHINE LEARNING AND BERT MODEL*.

Mohawesh, R., Xu, S., Tran, S.N., Ollington, R., Springer, M., Jararweh, Y. and Maqsood, S., (2021) *Fake Reviews Detection: A Survey*. *IEEE Access*, .

Mukherjee, A., Venkataraman, V., Liu, B. and Glance, N., (2013) *Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews*.

Poonguzhali, R., Sowmiya, S.F., Surendar, P. and Vasikaran, M., (2022) Fake Reviews Detection using Support Vector Machine. In: *International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2022 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp.1509–1512.

Rakibul, H. and Md. Rabiul, I., (2019) *Detection of fake online reviews using semi-supervised and supervised learning*. IEEE.

Reimers, N. and Gurevych, I., (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. [online] Available at: http://arxiv.org/abs/1908.10084.

Salunkhe, A., (2021) Attention-based Bidirectional LSTM for Deceptive Opinion Spam Classification. [online] Available at: http://arxiv.org/abs/2112.14789.

Sasikala, C., Ramya, S., RajeshKumar, S., Ramachandramoorthy, K.B., VijayaKumar, S. and Umapathi, K., (2023) Fake Review Detection and Classification Using Improved Convolutional Neural Network on Amazon Dataset. In: *Proceedings - 2023 3rd International Conference on Pervasive Computing and Social Networking, ICPCSN 2023*. Institute of Electrical and Electronics Engineers Inc., pp.398–403.

Saumya, S. and Singh, J.P., (2022) Spam review detection using LSTM autoencoder: an unsupervised approach. *Electronic Commerce Research*, 221, pp.113–133.

Shunxiang, Z., Aoqiang, Z., Guangli, Z., Zhongliang, W. and KuanChing, L., (2023) Building Fake Review Detection Model Based on Sentiment Intensity and PU Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Singh, D., Memoria, M. and Kumar, R., (2023) Deep Learning Based Model for Fake Review Detection. In: *2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT 2023*. Institute of Electrical and Electronics Engineers Inc., pp.92–95.

Singhal, R. and Kashef, R., (2023) A Weighted Stacking Ensemble Model With Sampling for Fake Reviews Detection. *IEEE Transactions on Computational Social Systems*.

Swathi, P., Raj, M.G.P., Babu, A. and Haritha, M., (2023) *Using Semi-Supervised and Supervised Learning for Fake Online Review Detection*. [online] *JETIR2303458 Journal of Emerging Technologies and Innovative Research*, JETIR. Available at: www.jetir.org.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., (2011) *Lexicon-Based Methods for Sentiment Analysis*.

Tang, H. and Cao, H., (2020) A review of research on detection of fake commodity reviews. In: *Journal of Physics: Conference Series*. IOP Publishing Ltd.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., (2017) Attention Is All You Need. [online] Available at: http://arxiv.org/abs/1706.03762.

Wang, C.C., Day, M.Y., Chen, C.C. and Liou, J.W., (2018) Detecting spamming reviews using long short-term memory recurrent neural network framework. In: *ACM International Conference Proceeding Series*. Association for Computing Machinery.

Wang Junren and Chen, J. and Z.W., (2023) A Novel Approach for Fake Review Detection Based on Reviewing Behavior and BERT Fused with Cosine Similarity. In: V.-N. and T.X. and W.J. Chen Jian and Huynh, ed., *Knowledge and Systems Sciences*. Singapore: Springer Nature Singapore, pp.18–32.

Yadav, S., Dharmela, Dr.G. and Mistry, K., (2021) *Fake Review Detection Using Machine Learning Techniques*. [online] Available at: www.jetir.org.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z. and Yu, P.S., (2018) TI-CNN: Convolutional Neural Networks for Fake News Detection. [online] Available at: http://arxiv.org/abs/1806.00749.

Zhang, D., Li, W., Niu, B. and Wu, C., (2023) A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decision Support Systems*, 166.

Zhong, M., Li, Z., Liu, S., Yang, B., Tan, R. and Qu, X., (2021) Fast Detection of Deceptive Reviews by Combining the Time Series and Machine Learning. *Complexity*, 2021.

Zou, X., Hu, Y., Tian, Z. and Shen, K., (2019) Logistic Regression Model Optimization and Case Analysis. In: *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*. pp.135–139.

APPENDIX A: RESEARCH PLAN

Table 6 Research Plan for the Thesis Fake Review Detection

Sl No	Task	Nov-23		Dec-23				Jan-24				Feb-24				Mar-24
		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15

1	Literature Review & Research Proposal	🟩	🟩													
2	Data Collection and Pre-processing			🟩	🟩											
2.1	**Data Cleanup**				🟩											
2.2	**Word2wec**				🟩											
3	Model Training					🟩	🟩	🟩	🟩	🟩						
3.1	**Logistic Regression**					🟩	🟩									
3.2	**SVM**					🟩	🟩									
3.3	**Naïve-Bayes Algorithm**					🟩	🟩									
3.5	**RNN-LSTM**							🟩	🟩	🟩						
3.6	**BERT**							🟩	🟩	🟩						
3.7	**Adversarial Learning**							🟩	🟩	🟩						
4	Model Evaluation								🟩	🟩	🟩	🟩				
5	Thesis Writing									🟩	🟩	🟩	🟩	🟩	🟩	🟩
5.1	Mid Thesis Submission										🟩	🟩				
5.2	Thesis Revision												🟩	🟩		
5.3	Final Thesis Submission														🟩	🟩

APPENDIX B: RESEARCH PROPOSAL

UNVEILING ONLINE DECEPTION: A DEEP LEARNING APPROACH TO DETECT FAKE REVIEWS

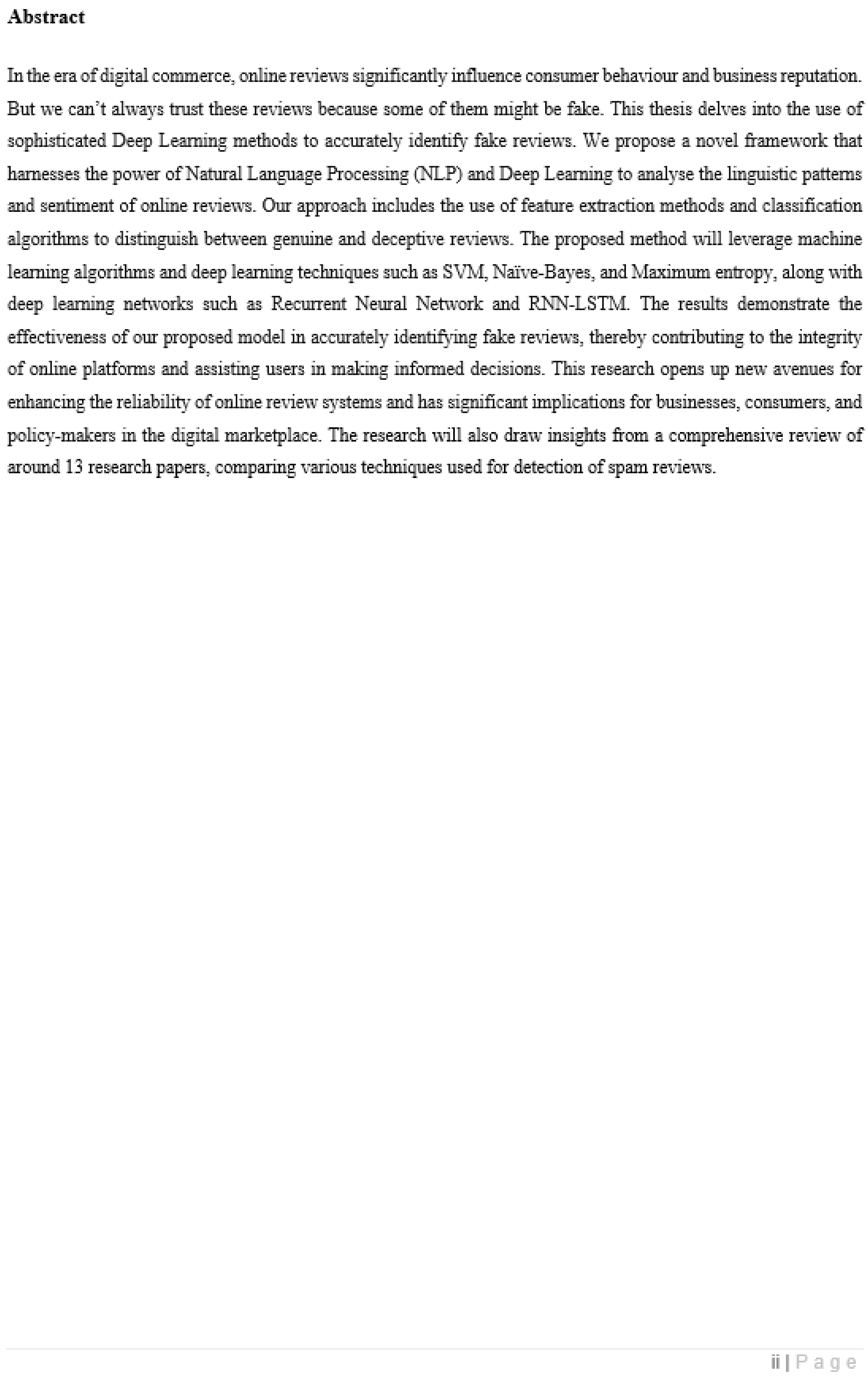
Masters in Data Science

Research Proposal

Liverpool John Moore University

Sridhar Pai Tonse

November 2023



Abstract

In the era of digital commerce, online reviews significantly influence consumer behaviour and business reputation. But we can’t always trust these reviews because some of them might be fake. This thesis delves into the use of sophisticated Deep Learning methods to accurately identify fake reviews. We propose a novel framework that harnesses the power of Natural Language Processing (NLP) and Deep Learning to analyse the linguistic patterns and sentiment of online reviews. Our approach includes the use of feature extraction methods and classification algorithms to distinguish between genuine and deceptive reviews. The proposed method will leverage machine learning algorithms and deep learning techniques such as SVM, Naïve-Bayes, and Maximum entropy, along with deep learning networks such as Recurrent Neural Network and RNN-LSTM. The results demonstrate the effectiveness of our proposed model in accurately identifying fake reviews, thereby contributing to the integrity of online platforms and assisting users in making informed decisions. This research opens up new avenues for enhancing the reliability of online review systems and has significant implications for businesses, consumers, and policy-makers in the digital marketplace. The research will also draw insights from a comprehensive review of around 13 research papers, comparing various techniques used for detection of spam reviews.

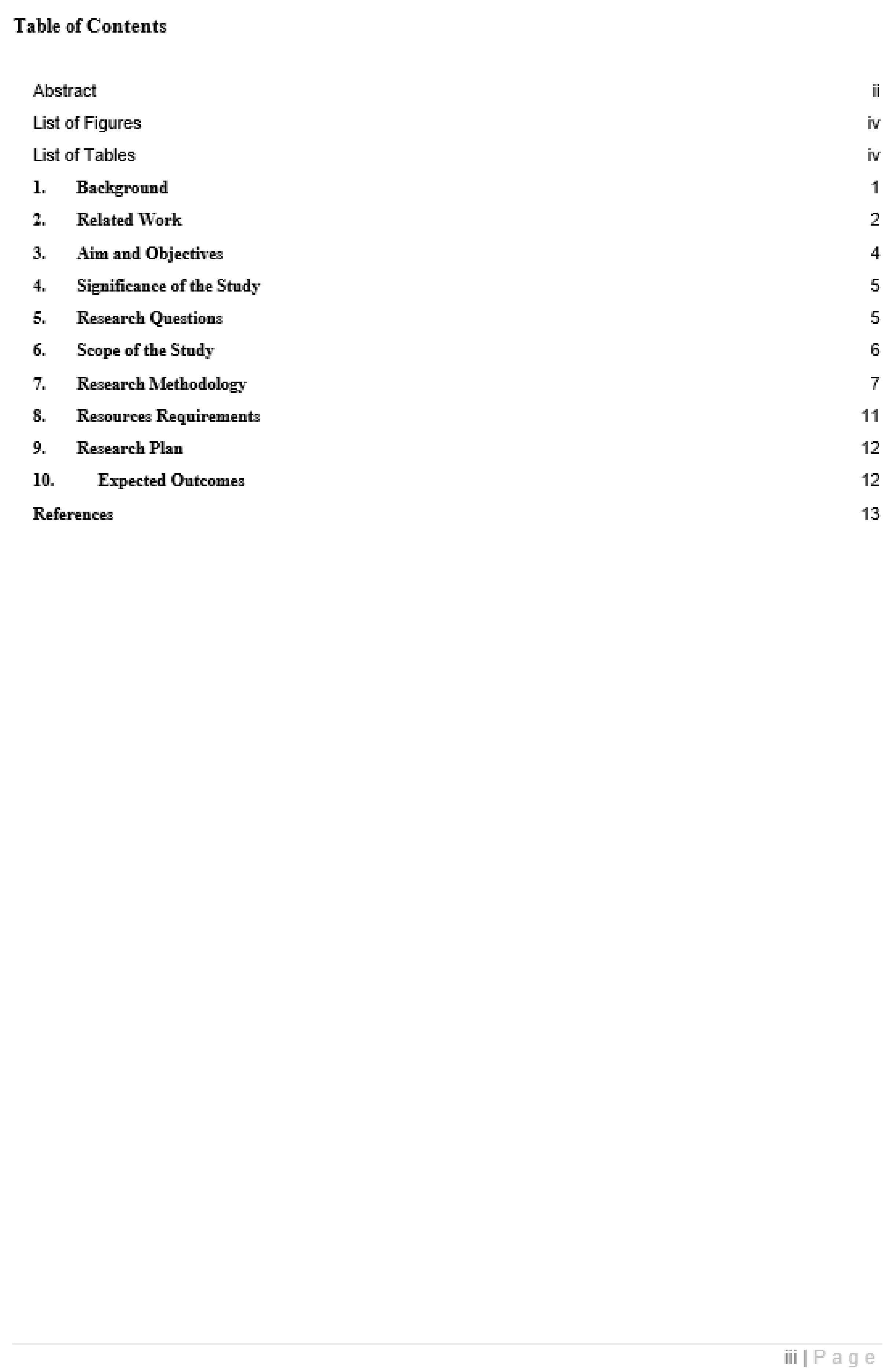


Table of Contents

Abstract ii

List of Figures iv

List of Tables iv

1. Background 1

2. Related Work 2

3. Aim and Objectives 4

4. Significance of the Study 5

5. Research Questions 5

6. Scope of the Study 6

7. Research Methodology 7

8. Resources Requirements 11

9. Research Plan 12

10. Expected Outcomes 12

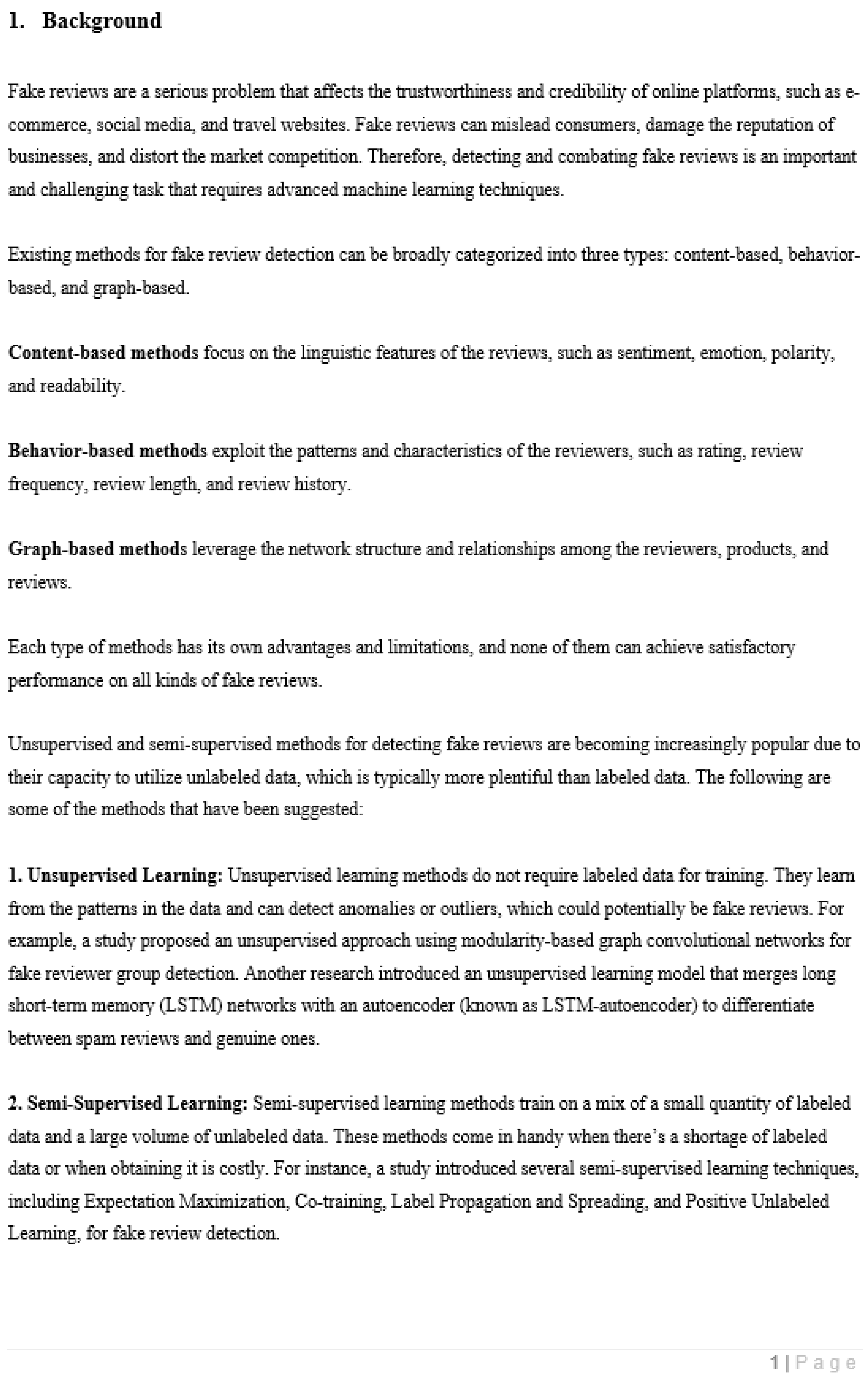
References 13

List of Figures

Figure 1 Block Diagram for Fake review detection methodology 11
Figure 2 An illustration of online review texts described by CNN-LSTM model 12
Figure 3 RNN – LSTM model 12

List of Tables

Table 1 Timeline for Thesis (Gantt Chart) 16



1. Background

Fake reviews are a serious problem that affects the trustworthiness and credibility of online platforms, such as e-commerce, social media, and travel websites. Fake reviews can mislead consumers, damage the reputation of businesses, and distort the market competition. Therefore, detecting and combating fake reviews is an important and challenging task that requires advanced machine learning techniques.

Existing methods for fake review detection can be broadly categorized into three types: content-based, behavior-based, and graph-based.

Content-based methods focus on the linguistic features of the reviews, such as sentiment, emotion, polarity, and readability.

Behavior-based methods exploit the patterns and characteristics of the reviewers, such as rating, review frequency, review length, and review history.

Graph-based methods leverage the network structure and relationships among the reviewers, products, and reviews.

Each type of methods has its own advantages and limitations, and none of them can achieve satisfactory performance on all kinds of fake reviews.

Unsupervised and semi-supervised methods for detecting fake reviews are becoming increasingly popular due to their capacity to utilize unlabeled data, which is typically more plentiful than labeled data. The following are some of the methods that have been suggested:

1. Unsupervised Learning: Unsupervised learning methods do not require labeled data for training. They learn from the patterns in the data and can detect anomalies or outliers, which could potentially be fake reviews. For example, a study proposed an unsupervised approach using modularity-based graph convolutional networks for fake reviewer group detection. Another research introduced an unsupervised learning model that merges long short-term memory (LSTM) networks with an autoencoder (known as LSTM-autoencoder) to differentiate between spam reviews and genuine ones.

2. Semi-Supervised Learning: Semi-supervised learning methods train on a mix of a small quantity of labeled data and a large volume of unlabeled data. These methods come in handy when there's a shortage of labeled data or when obtaining it is costly. For instance, a study introduced several semi-supervised learning techniques, including Expectation Maximization, Co-training, Label Propagation and Spreading, and Positive Unlabeled Learning, for fake review detection.

These methods offer promising results in detecting fake reviews. Nevertheless, they present challenges, including the complexity of validating results due to the absence of a clear standard for comparison and vulnerability to the dynamic tactics employed by spammers. Hence, ongoing research and development efforts are crucial to enhance these approaches and address the ever-changing landscape of fake reviews.

Research Gap: The majority of current methods for detecting fake reviews rely on supervised learning, which necessitates the use of labeled data for both training and testing. However, obtaining labeled data can be challenging as it is often limited in quantity, potentially noisy, and specific to certain domains, which restricts the ability of supervised methods to generalize and scale. Furthermore, supervised methods may struggle to identify the nuanced and changing characteristics of fake reviews, which are constantly evolving to evade detection. As a result, there is a demand for the development of unsupervised or semi-supervised methods that can detect fake reviews without the need for labeled data or with minimal human intervention.

Purpose of this Study is to study various machine learning and deep learning techniques used for fake review detection and come up best possible method that can be applied using unsupervised methods or semi supervised methods for unlabeled data or with minimal human intervention.

2. Related Work

(Yadav et al., 2021), defines fake reviews as untruthful reviews that are written to promote or defame a target object. It discusses the use of machine learning techniques for detecting fake reviews. The paper reports that Support Vector Machine achieves the highest accuracy of 93.33%, followed by KStar with 91.67%, K-Nearest Neighbours with 90%, Decision Tree with 88.33%, and Naïve Bayes with 86.67%.

(Jain et al., 2021), The authors propose employing a fully supervised approach for detecting opinion spammers in online reviews. They leverage labelled data to distinguish between genuine and fraudulent reviews, employing various machine learning algorithms for classification on datasets containing Yelp hotel and restaurant reviews. The classification is conducted on a dataset enriched with engineered features. As per the authors, Logistic Regression consistently outperforms other algorithms in the majority of cases.

(Zhang et al., 2023), The paper introduces an innovative, comprehensive framework that integrates behavioural and textual data to identify fake reviewers. This framework is tested on two authentic datasets from Yelp.com, and its performance is compared with that of the most advanced existing models. The results indicate that the proposed framework outperforms the others in detecting fake reviewers.

(Duma et al., 2023), The paper presents a model that merges a bidirectional encoder representation from transformers (BERT) model with a convolutional neural network (CNN) - long short-term memory (LSTM) model to learn hidden text feature vectors. The study reveals that the suggested model achieves the highest precision of 95.83%. This is followed by a detection of spam reviews through a hierarchical attention architecture model with

94.17%, a behavioural feature generative adversarial network (bfGAN) model with 93.33%, and a feature-aware bidirectional convolutional LSTM (FABC) model with 92.5%.

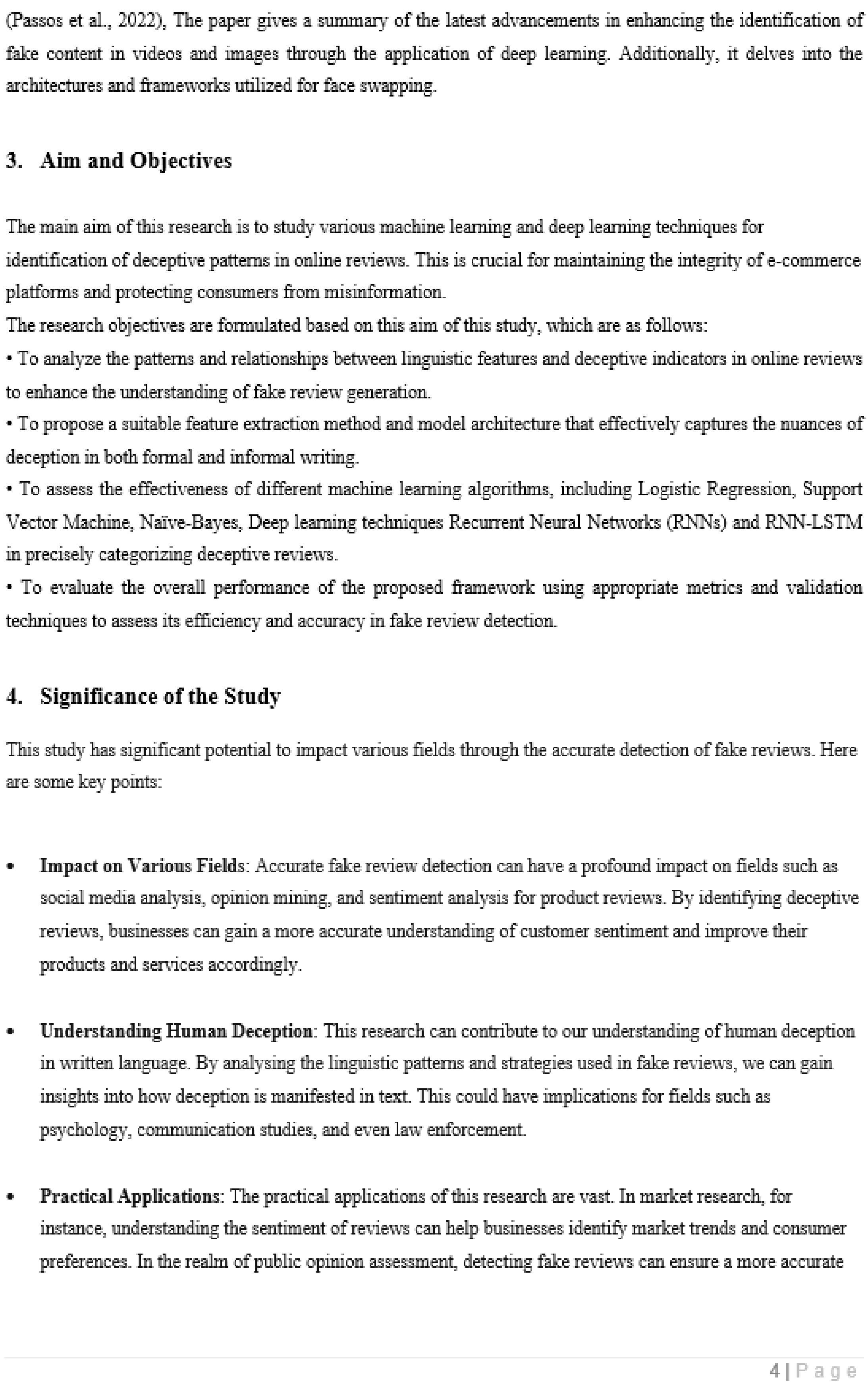
(Shunxiang et al., 2023), The paper proposes a model called SIPUL, which uses sentiment intensity and PU learning to detect fake reviews from streaming data. PU learning is a semi-supervised learning method that only uses positive and unlabelled samples. The assessment of the model indicates that it surpasses the baseline models in terms of accuracy, precision, recall, and F1-score. It demonstrates a strong capability to effectively identify deceptive fake reviews.

(Poonguzhali et al., 2022), the paper uses SVM to classify the reviews into positive, negative, and neutral groups based on the text and ratings. Then, it predicts the fake reviews based on the user id and booking id. The model results show the model achieves an accuracy of 92.5% for fake reviews detection and 89.6% for product recommendation.

In "Unmasking Deception (Deshai and Rao, 2023), proposes two pronged approach. a) Harnessing power of CNN by scrutinizing both text of the review and additional metadata such as reviewer history and rating patterns. This provides a richer picture for analysing intent and uncovering deception. b) Adaptive PSO (particle swarm optimization)'s intelligent guidance. The Results: The hybrid system shines with exceptional performance ie 99.4% accuracy and Generalizability across platforms such as tech reviews on Amazon to travel experiences on TripAdvisor, this system adapts to diverse datasets, proving its real-world effectiveness. Deshai acknowledges the ever-evolving nature of fake reviews and suggests further research directions, including utilizing advanced attention mechanisms like Bi-LSTM and BERT for even deeper analysis.

(Singhal and Kashef, 2023) introduces a novel multi-model architecture, leveraging both the collective intelligence of diverse machine learning algorithms and a targeted sampling strategy, to achieve superior performance in fake review detection. proposes a weighted stacking ensemble model, where individual base learners, including logistic regression and support vector machines, collaborate to form a more robust and accurate classifier. This ensemble's power lies in its heterogeneity and dynamic weighting, which assigns greater influence to base learners demonstrably adept at discerning genuine from deceptive reviews. This approach transcends limitations inherent in singular models, harnessing the combined analytical strength of multiple perspectives.

To further optimize performance, Singhal employs a strategic sampling technique. By focusing on reviews exhibiting characteristics indicative of deception, the model efficiently allocates its analytical resources, prioritizing those most likely to harbour hidden falsehoods. This targeted approach significantly improves detection accuracy while minimizing computational overhead. The empirical results showcase the efficacy of Singhal's framework. The proposed model outperforms baseline approaches, achieving a demonstrably higher degree of precision in identifying fake reviews. This signifies a significant advancement in the fight against online deception, empowering both consumers and businesses to operate with greater confidence in the online marketplace.



(Passos et al., 2022), The paper gives a summary of the latest advancements in enhancing the identification of fake content in videos and images through the application of deep learning. Additionally, it delves into the architectures and frameworks utilized for face swapping.

3. Aim and Objectives

The main aim of this research is to study various machine learning and deep learning techniques for identification of deceptive patterns in online reviews. This is crucial for maintaining the integrity of e-commerce platforms and protecting consumers from misinformation.

The research objectives are formulated based on this aim of this study, which are as follows:

• To analyze the patterns and relationships between linguistic features and deceptive indicators in online reviews to enhance the understanding of fake review generation.

• To propose a suitable feature extraction method and model architecture that effectively captures the nuances of deception in both formal and informal writing.

• To assess the effectiveness of different machine learning algorithms, including Logistic Regression, Support Vector Machine, Naïve-Bayes, Deep learning techniques Recurrent Neural Networks (RNNs) and RNN-LSTM in precisely categorizing deceptive reviews.

• To evaluate the overall performance of the proposed framework using appropriate metrics and validation techniques to assess its efficiency and accuracy in fake review detection.

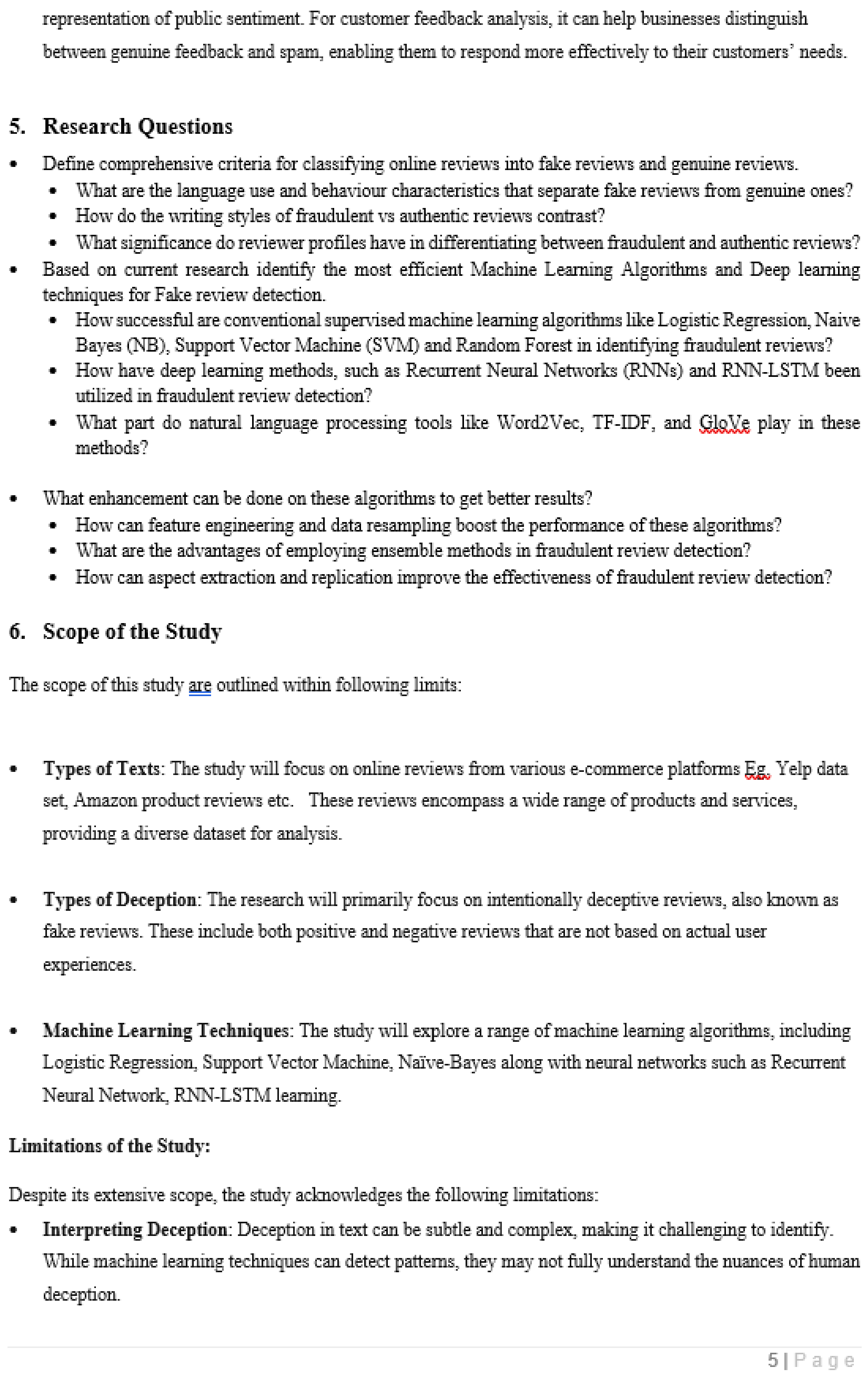
4. Significance of the Study

This study has significant potential to impact various fields through the accurate detection of fake reviews. Here are some key points:

- **Impact on Various Fields**: Accurate fake review detection can have a profound impact on fields such as social media analysis, opinion mining, and sentiment analysis for product reviews. By identifying deceptive reviews, businesses can gain a more accurate understanding of customer sentiment and improve their products and services accordingly.

- **Understanding Human Deception**: This research can contribute to our understanding of human deception in written language. By analysing the linguistic patterns and strategies used in fake reviews, we can gain insights into how deception is manifested in text. This could have implications for fields such as psychology, communication studies, and even law enforcement.

- **Practical Applications**: The practical applications of this research are vast. In market research, for instance, understanding the sentiment of reviews can help businesses identify market trends and consumer preferences. In the realm of public opinion assessment, detecting fake reviews can ensure a more accurate



representation of public sentiment. For customer feedback analysis, it can help businesses distinguish between genuine feedback and spam, enabling them to respond more effectively to their customers' needs.

5. Research Questions

- Define comprehensive criteria for classifying online reviews into fake reviews and genuine reviews.
  - What are the language use and behaviour characteristics that separate fake reviews from genuine ones?
  - How do the writing styles of fraudulent vs authentic reviews contrast?
  - What significance do reviewer profiles have in differentiating between fraudulent and authentic reviews?
- Based on current research identify the most efficient Machine Learning Algorithms and Deep learning techniques for Fake review detection.
  - How successful are conventional supervised machine learning algorithms like Logistic Regression, Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest in identifying fraudulent reviews?
  - How have deep learning methods, such as Recurrent Neural Networks (RNNs) and RNN-LSTM been utilized in fraudulent review detection?
  - What part do natural language processing tools like Word2Vec, TF-IDF, and GloVe play in these methods?

- What enhancement can be done on these algorithms to get better results?
  - How can feature engineering and data resampling boost the performance of these algorithms?
  - What are the advantages of employing ensemble methods in fraudulent review detection?
  - How can aspect extraction and replication improve the effectiveness of fraudulent review detection?

6. Scope of the Study

The scope of this study are outlined within following limits:

- **Types of Texts**: The study will focus on online reviews from various e-commerce platforms Eg. Yelp data set, Amazon product reviews etc. These reviews encompass a wide range of products and services, providing a diverse dataset for analysis.

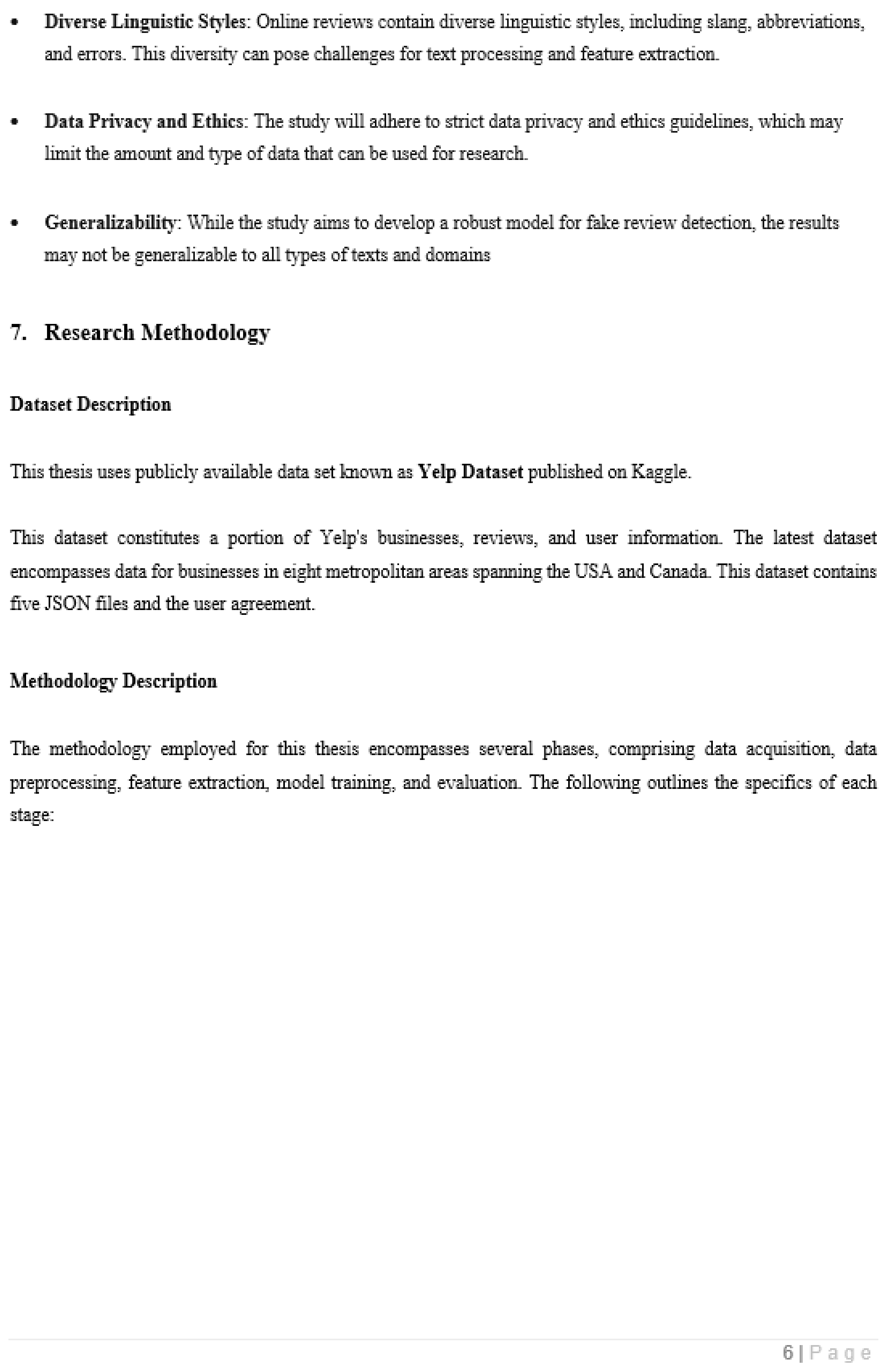
- **Types of Deception**: The research will primarily focus on intentionally deceptive reviews, also known as fake reviews. These include both positive and negative reviews that are not based on actual user experiences.

- **Machine Learning Techniques**: The study will explore a range of machine learning algorithms, including Logistic Regression, Support Vector Machine, Naïve-Bayes along with neural networks such as Recurrent Neural Network, RNN-LSTM learning.

Limitations of the Study:

Despite its extensive scope, the study acknowledges the following limitations:

- **Interpreting Deception**: Deception in text can be subtle and complex, making it challenging to identify. While machine learning techniques can detect patterns, they may not fully understand the nuances of human deception.



- **Diverse Linguistic Styles**: Online reviews contain diverse linguistic styles, including slang, abbreviations, and errors. This diversity can pose challenges for text processing and feature extraction.

- **Data Privacy and Ethics**: The study will adhere to strict data privacy and ethics guidelines, which may limit the amount and type of data that can be used for research.

- **Generalizability**: While the study aims to develop a robust model for fake review detection, the results may not be generalizable to all types of texts and domains

7. Research Methodology

Dataset Description

This thesis uses publicly available data set known as **Yelp Dataset** published on Kaggle.

This dataset constitutes a portion of Yelp's businesses, reviews, and user information. The latest dataset encompasses data for businesses in eight metropolitan areas spanning the USA and Canada. This dataset contains five JSON files and the user agreement.

Methodology Description

The methodology employed for this thesis encompasses several phases, comprising data acquisition, data preprocessing, feature extraction, model training, and evaluation. The following outlines the specifics of each stage:

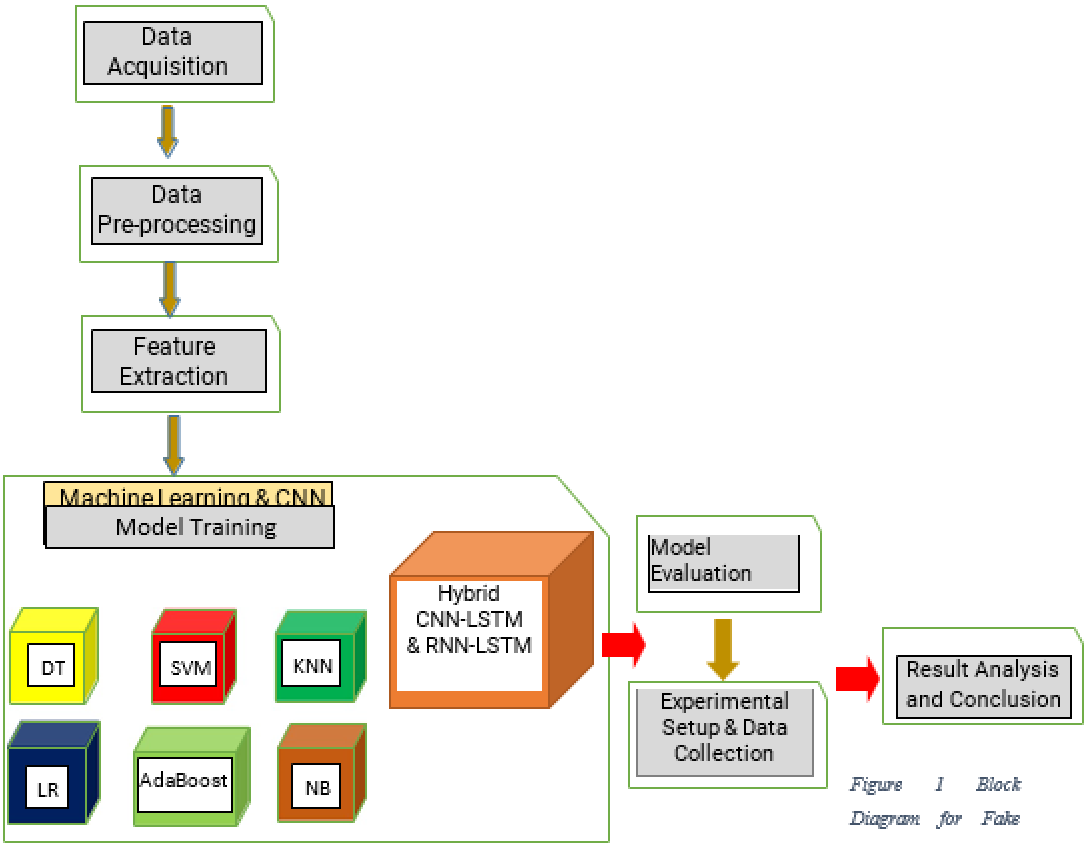


Figure 1 Block Diagram for Fake review detection methodology

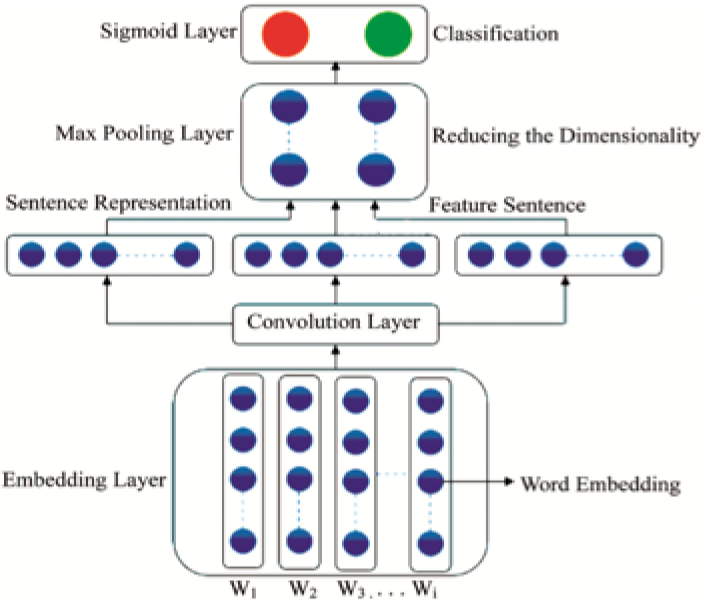


Figure 2 An illustration of online review texts described by CNN-LSTM model

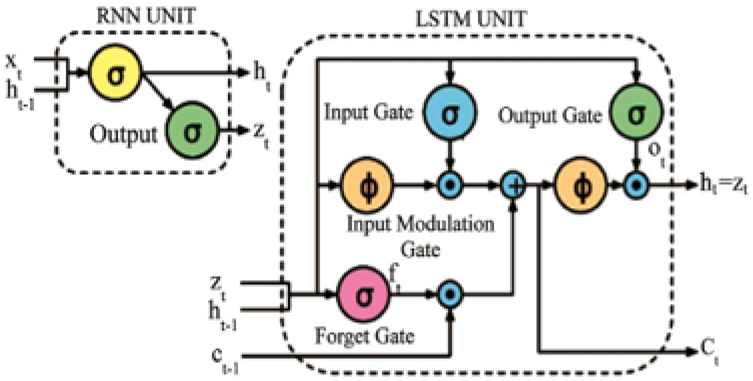


Figure 3 RNN – LSTM model

1. **Data Acquisition:** For our investigation, we utilized the Yelp dataset focusing on hotels and restaurants. Yelp provides various attributes related to product or service reviews and user-related data.

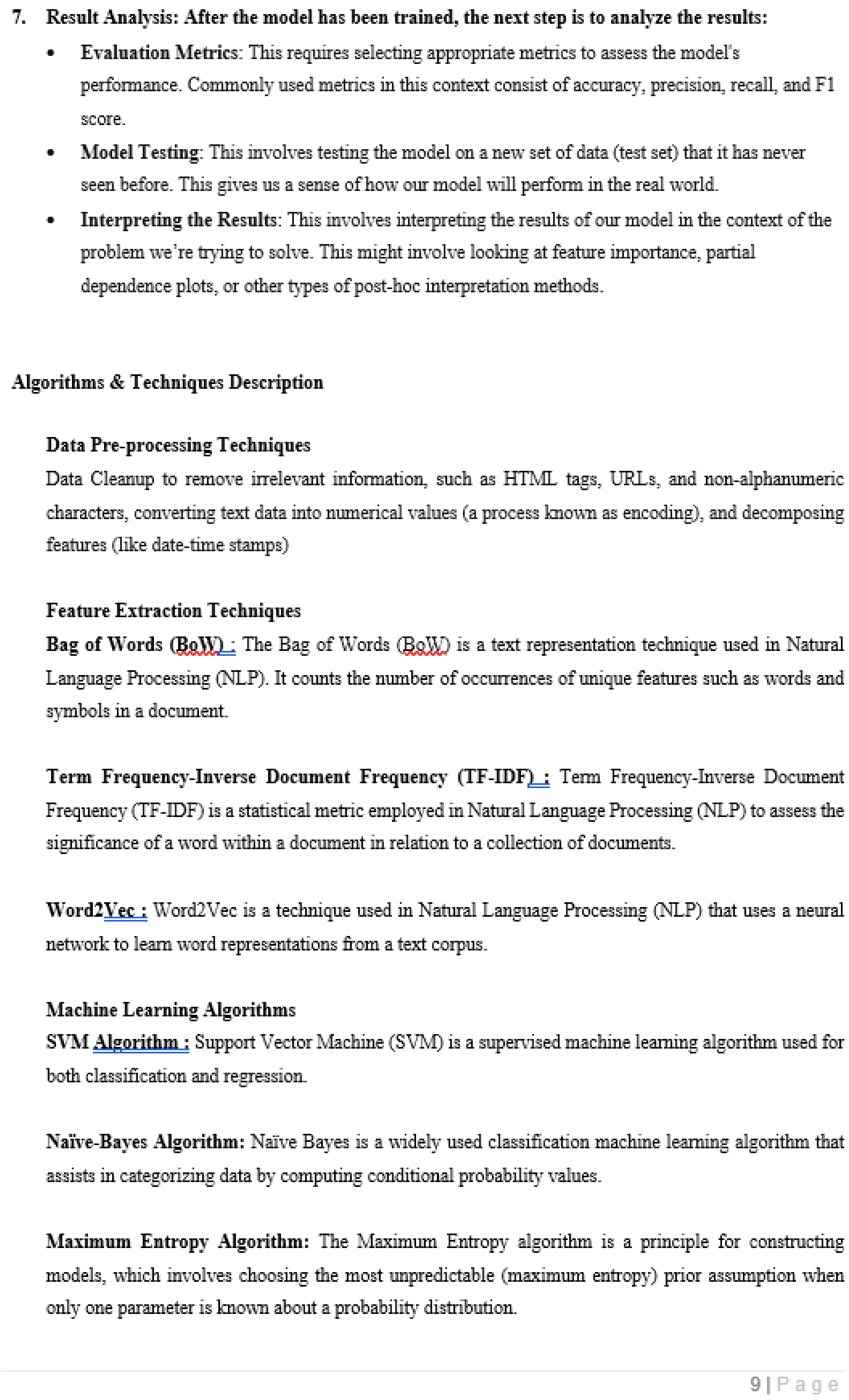
2. **Data Preprocessing:** Next step in our methodology will be data preprocessing. The data will be cleaned to remove any irrelevant information, such as HTML tags, URLs, and non-alphanumeric characters. N-Grams, Normalization, Stemming, Lemmatization, Vectorization, Word2Vec, One hot encoding techniques will be applied in preprocessing to prepare for Machine Learning algorithms along with Neural networks.

3. **Feature Extraction:** The next step will be feature extraction. We will use various techniques to extract features from the preprocessed data. These may include Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec. We may also explore the use of pre-trained models like BERT for feature extraction.

4. **Model Training:** After extracting the features, we will proceed to train diverse machine learning models on the dataset. These models encompass SVM, Naïve-Bayes, and maximum entropy, as well as neural networks like recurrent neural network, CNN-LSTM, and LSTM-RNN. The objective is to train the models to classify reviews into either genuine or fake categories.

5. **Model Evaluation:** Finally, we will evaluate the performance of the models using appropriate metrics. These may include accuracy, precision, recall, and F1 score. We will also use techniques such as cross-validation to ensure the robustness of our results.

6. **Experimental Setup and Data Collection Procedures:** The experimental setup will involve setting up a suitable computing environment for running the machine learning models. This may involve the use of high-performance computing clusters or cloud-based platforms. The data collection procedures will involve scraping online reviews from various e-commerce websites and manually labeling a subset of these reviews as either genuine or fake.



7. **Result Analysis: After the model has been trained, the next step is to analyze the results:**
   - **Evaluation Metrics**: This requires selecting appropriate metrics to assess the model's performance. Commonly used metrics in this context consist of accuracy, precision, recall, and F1 score.
   - **Model Testing**: This involves testing the model on a new set of data (test set) that it has never seen before. This gives us a sense of how our model will perform in the real world.
   - **Interpreting the Results**: This involves interpreting the results of our model in the context of the problem we're trying to solve. This might involve looking at feature importance, partial dependence plots, or other types of post-hoc interpretation methods.

Algorithms & Techniques Description

Data Pre-processing Techniques

Data Cleanup to remove irrelevant information, such as HTML tags, URLs, and non-alphanumeric characters, converting text data into numerical values (a process known as encoding), and decomposing features (like date-time stamps)

Feature Extraction Techniques

Bag of Words (BoW) : The Bag of Words (BoW) is a text representation technique used in Natural Language Processing (NLP). It counts the number of occurrences of unique features such as words and symbols in a document.

Term Frequency-Inverse Document Frequency (TF-IDF) : Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical metric employed in Natural Language Processing (NLP) to assess the significance of a word within a document in relation to a collection of documents.

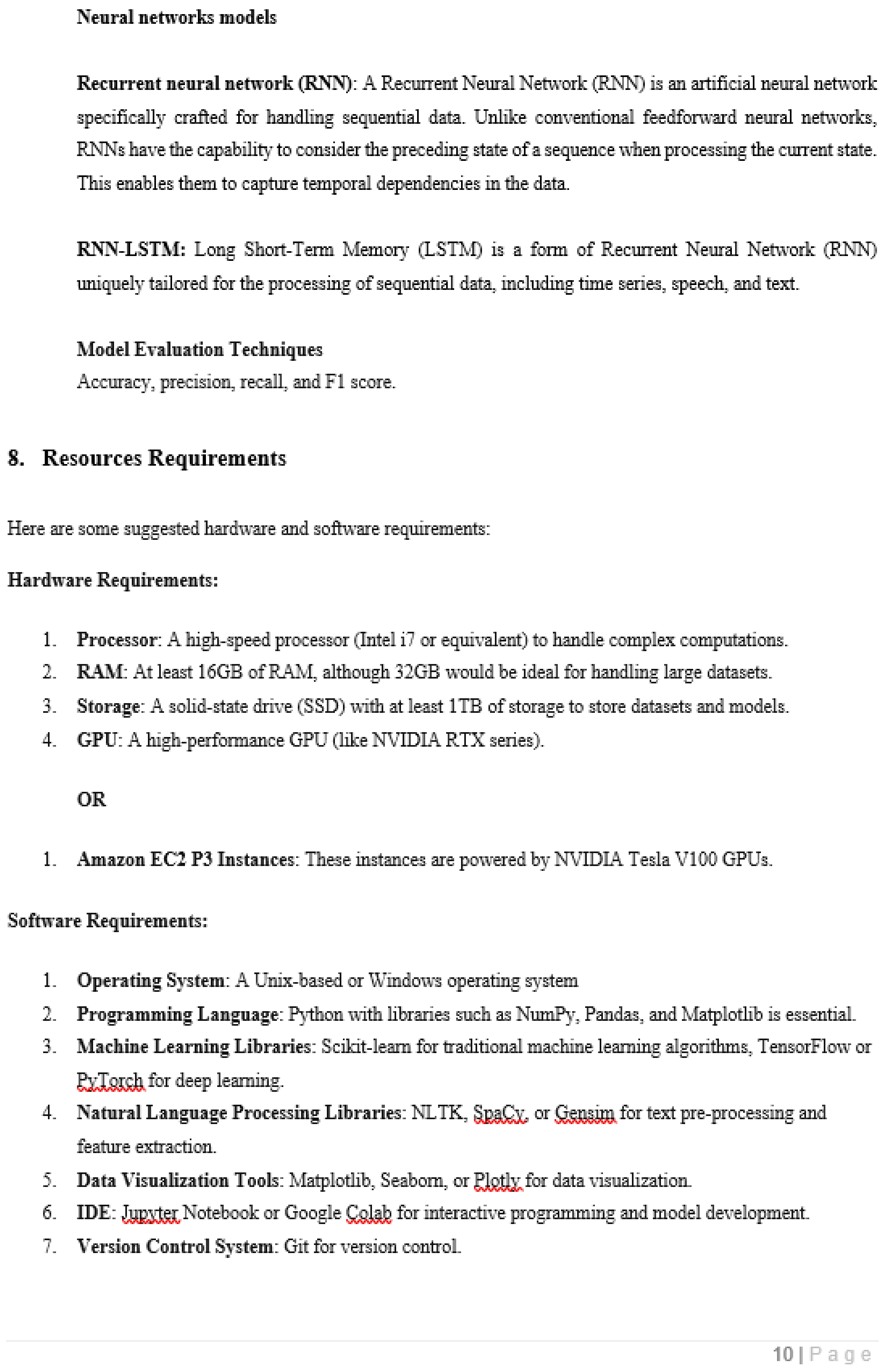
Word2Vec : Word2Vec is a technique used in Natural Language Processing (NLP) that uses a neural network to learn word representations from a text corpus.

Machine Learning Algorithms

SVM Algorithm : Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression.

Naïve-Bayes Algorithm: Naïve Bayes is a widely used classification machine learning algorithm that assists in categorizing data by computing conditional probability values.

Maximum Entropy Algorithm: The Maximum Entropy algorithm is a principle for constructing models, which involves choosing the most unpredictable (maximum entropy) prior assumption when only one parameter is known about a probability distribution.



Neural networks models

Recurrent neural network (RNN): A Recurrent Neural Network (RNN) is an artificial neural network specifically crafted for handling sequential data. Unlike conventional feedforward neural networks, RNNs have the capability to consider the preceding state of a sequence when processing the current state. This enables them to capture temporal dependencies in the data.

RNN-LSTM: Long Short-Term Memory (LSTM) is a form of Recurrent Neural Network (RNN) uniquely tailored for the processing of sequential data, including time series, speech, and text.

Model Evaluation Techniques
Accuracy, precision, recall, and F1 score.

8. Resources Requirements

Here are some suggested hardware and software requirements:

Hardware Requirements:

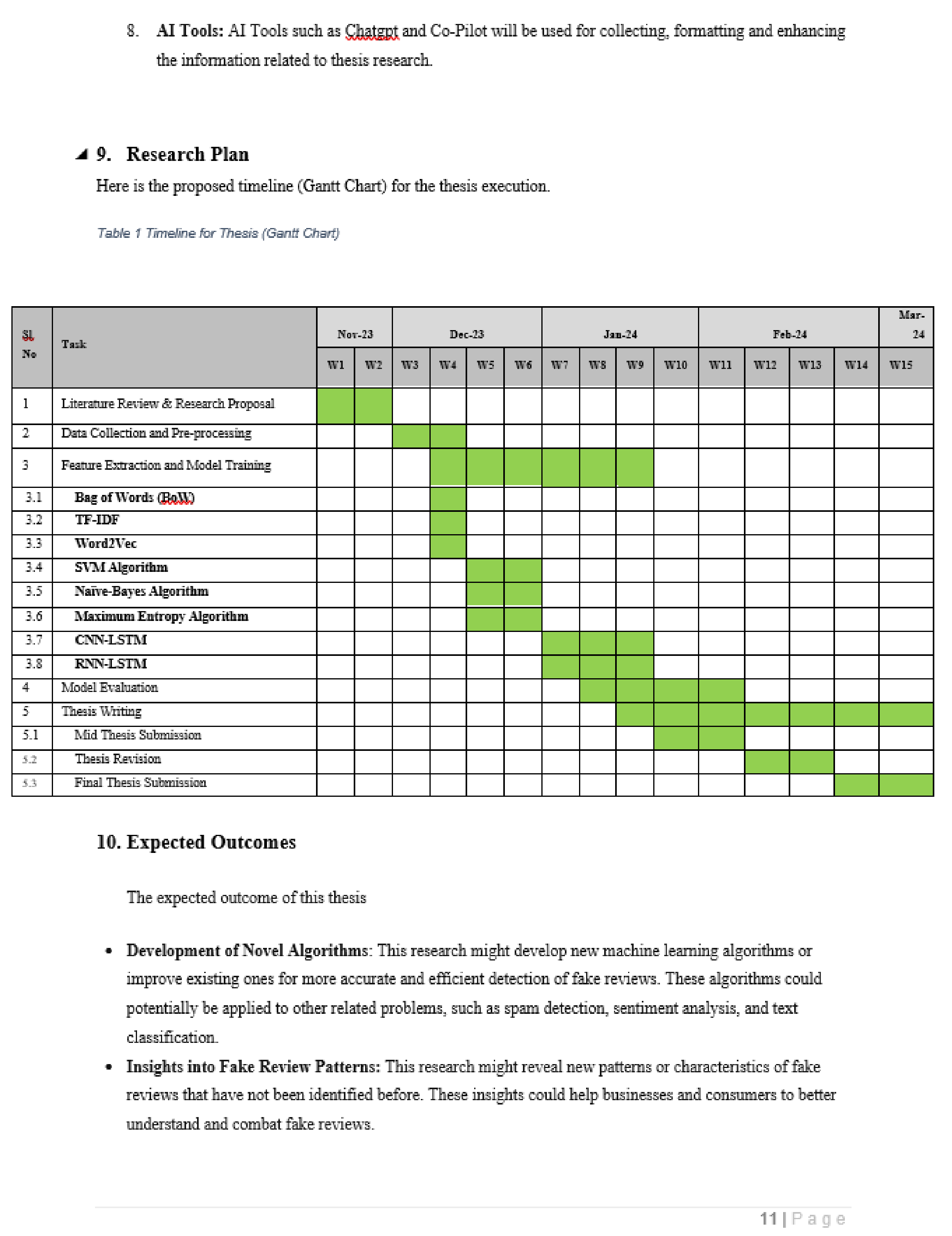
1. **Processor**: A high-speed processor (Intel i7 or equivalent) to handle complex computations.
2. **RAM**: At least 16GB of RAM, although 32GB would be ideal for handling large datasets.
3. **Storage**: A solid-state drive (SSD) with at least 1TB of storage to store datasets and models.
4. **GPU**: A high-performance GPU (like NVIDIA RTX series).

OR

1. **Amazon EC2 P3 Instances**: These instances are powered by NVIDIA Tesla V100 GPUs.

Software Requirements:

1. **Operating System**: A Unix-based or Windows operating system
2. **Programming Language**: Python with libraries such as NumPy, Pandas, and Matplotlib is essential.
3. **Machine Learning Libraries**: Scikit-learn for traditional machine learning algorithms, TensorFlow or PyTorch for deep learning.
4. **Natural Language Processing Libraries**: NLTK, SpaCy, or Gensim for text pre-processing and feature extraction.
5. **Data Visualization Tools**: Matplotlib, Seaborn, or Plotly for data visualization.
6. **IDE**: Jupyter Notebook or Google Colab for interactive programming and model development.
7. **Version Control System**: Git for version control.



8. **AI Tools:** AI Tools such as Chatgpt and Co-Pilot will be used for collecting, formatting and enhancing the information related to thesis research.

9. Research Plan

Here is the proposed timeline (Gantt Chart) for the thesis execution.

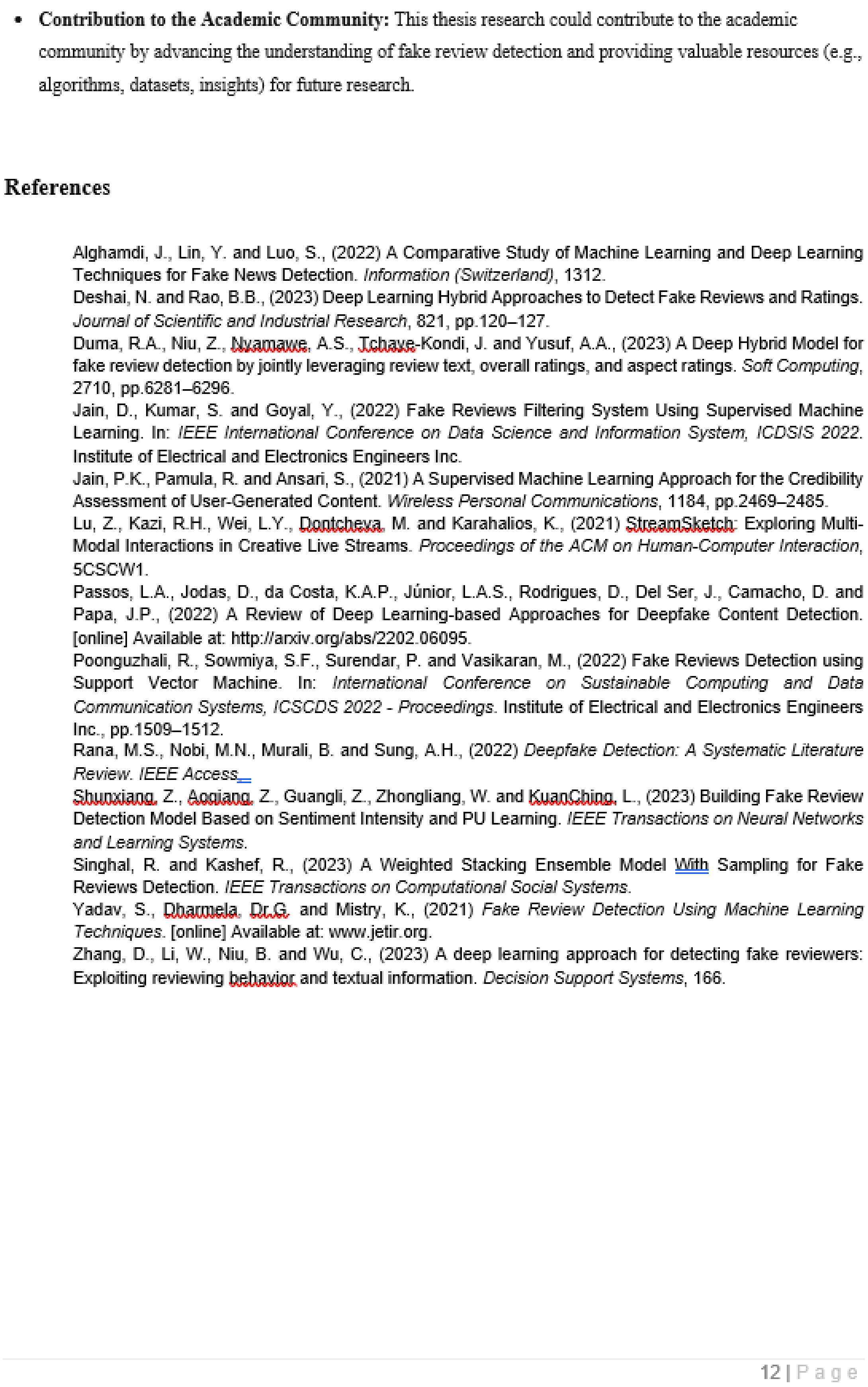
Table 1 Timeline for Thesis (Gantt Chart)

SL No	Task	Nov-23		Dec-23				Jan-24				Feb-24				Mar-24
		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15
1	Literature Review & Research Proposal	🟩	🟩													
2	Data Collection and Pre-processing			🟩	🟩											
3	Feature Extraction and Model Training				🟩	🟩	🟩	🟩	🟩	🟩						
3.1	**Bag of Words (BoW)**				🟩											
3.2	**TF-IDF**				🟩											
3.3	**Word2Vec**				🟩											
3.4	**SVM Algorithm**					🟩	🟩									
3.5	**Naïve-Bayes Algorithm**					🟩	🟩									
3.6	**Maximum Entropy Algorithm**					🟩	🟩									
3.7	**CNN-LSTM**							🟩	🟩	🟩						
3.8	**RNN-LSTM**							🟩	🟩	🟩						
4	Model Evaluation								🟩	🟩	🟩	🟩				
5	Thesis Writing									🟩	🟩	🟩	🟩	🟩	🟩	🟩
5.1	Mid Thesis Submission										🟩	🟩				
5.2	Thesis Revision												🟩	🟩		
5.3	Final Thesis Submission														🟩	🟩

10. Expected Outcomes

The expected outcome of this thesis

- **Development of Novel Algorithms**: This research might develop new machine learning algorithms or improve existing ones for more accurate and efficient detection of fake reviews. These algorithms could potentially be applied to other related problems, such as spam detection, sentiment analysis, and text classification.
- **Insights into Fake Review Patterns:** This research might reveal new patterns or characteristics of fake reviews that have not been identified before. These insights could help businesses and consumers to better understand and combat fake reviews.



- **Contribution to the Academic Community:** This thesis research could contribute to the academic community by advancing the understanding of fake review detection and providing valuable resources (e.g., algorithms, datasets, insights) for future research.

References

Alghamdi, J., Lin, Y. and Luo, S., (2022) A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection. *Information (Switzerland)*, 1312.

Deshai, N. and Rao, B.B., (2023) Deep Learning Hybrid Approaches to Detect Fake Reviews and Ratings. *Journal of Scientific and Industrial Research*, 821, pp.120–127.

Duma, R.A., Niu, Z., Nyamawe, A.S., Tchaye-Kondi, J. and Yusuf, A.A., (2023) A Deep Hybrid Model for fake review detection by jointly leveraging review text, overall ratings, and aspect ratings. *Soft Computing*, 2710, pp.6281–6296.

Jain, D., Kumar, S. and Goyal, Y., (2022) Fake Reviews Filtering System Using Supervised Machine Learning. In: *IEEE International Conference on Data Science and Information System, ICDSIS 2022*. Institute of Electrical and Electronics Engineers Inc.

Jain, P.K., Pamula, R. and Ansari, S., (2021) A Supervised Machine Learning Approach for the Credibility Assessment of User-Generated Content. *Wireless Personal Communications*, 1184, pp.2469–2485.

Lu, Z., Kazi, R.H., Wei, L.Y., Dontcheva, M. and Karahalios, K., (2021) StreamSketch: Exploring Multi-Modal Interactions in Creative Live Streams. *Proceedings of the ACM on Human-Computer Interaction*, 5CSCW1.

Passos, L.A., Jodas, D., da Costa, K.A.P., Júnior, L.A.S., Rodrigues, D., Del Ser, J., Camacho, D. and Papa, J.P., (2022) A Review of Deep Learning-based Approaches for Deepfake Content Detection. [online] Available at: http://arxiv.org/abs/2202.06095.

Poonguzhali, R., Sowmiya, S.F., Surendar, P. and Vasikaran, M., (2022) Fake Reviews Detection using Support Vector Machine. In: *International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2022 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp.1509–1512.

Rana, M.S., Nobi, M.N., Murali, B. and Sung, A.H., (2022) *Deepfake Detection: A Systematic Literature Review. IEEE Access*.

Shunxiang, Z., Aoqiang, Z., Guangli, Z., Zhongliang, W. and KuanChing, L., (2023) Building Fake Review Detection Model Based on Sentiment Intensity and PU Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Singhal, R. and Kashef, R., (2023) A Weighted Stacking Ensemble Model With Sampling for Fake Reviews Detection. *IEEE Transactions on Computational Social Systems*.

Yadav, S., Dharmela, Dr.G. and Mistry, K., (2021) *Fake Review Detection Using Machine Learning Techniques*. [online] Available at: www.jetir.org.

Zhang, D., Li, W., Niu, B. and Wu, C., (2023) A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decision Support Systems*, 166.

APPENDIX C: ETHICS FORMS