National College of Ireland

Online Review Classification using Machine Learning and Deep Learning Algorithms

MSc Research Project
Data Analytics

Akshaansh Gautam
Student ID: x20151438

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Akshaansh Gautam
Student ID:	x20151438
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	15/08/2022
Project Title:	Online Review Classification using Machine Learning and Deep Learning Algorithms
Word Count:	XXX
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Online Review Classification using Machine Learning and Deep Learning Algorithms

Akshaansh Gautam
x20151438

Abstract

Many online e-commerce platforms and websites that provide services online like amazon allows users to give their feedbacks for the items they have purchased. Majority of the online customers do further research about the product quality and experience before purchasing. The text data generated in the form of reviews can be analysed using sentiment analysis to obtain the sentiments of the users regarding the product and determine the fake reviews. The analysis will assist the marketers to grasp their customers preferences and prepare strategies that can satisfy the needs of customers and a well as the sellers. This report uses the best of machine learning algorithms like Logistic Regression, Decision Trees, Support Vector Machines, Naïve Bayes, Random Forest, XGBoost, Extra Trees with different feature extractions like CountVectorizer and TF-IDF Vectorizer. It also uses best of Deep Learning Sequence models like LSTM, Bi-Directional LSTM, LSTM with attention layers, GRU, Bi-Directional GRU, GRU with attention layers. It can be found that the best of the models for the original dataset is Bi-Directional LSTM with 93.75% accuracy followed by Bi-Directional GRU with 93.54% accuracy. While the ReviewNet concept has achieved an accuracy of 95.57% accuracy with SVM followed by 94.40% accuracy with linear SVC. Bi-Directional LSTM and GRU have got 93.86% and 93.35% respectively. ReviewNet achieved an accuracy of 95.57% accuracy in comparison to 93.75% for the original dataset.

1 Introduction

As many services like purchasing a product, booking a hotel, booking online tickets are now available online, customers have become heavily dependent on their services to fulfil their needs. Apart from selling their services these online platforms have also allowed their customers to give feedback/reviews about their experience. E-commerce platforms like amazon are preferred over other platforms as previous customers have given their detailed reviews/feedback on the items they have purchased before. This is very useful as customers have a habit of researching before purchasing a product. The reputation, success and profitability of any organisation, product or services is heavily dependent upon the reviews given by the previous user as its credibility can heavily affect the decision of the users. Due to this several cases of fake reviews being posted online are coming forward that are manipulating the decisions made by the users and are affecting the online platforms. The reviews not only help the customers but the seller as well. According to the reviews left by the users as they can change their strategy or improve the quality of their product.

As millions of reviews are posted on a short interval it becomes really difficult for the seller or the organization to identify the fake reviews from the real ones. Just like real reviews and ratings can help the platform and its users, fake reviews and rating however can harm their reputation as well. The cases of fake reviews are rising confusing the customer and increasing difficulties to reach decision about a product. Users who are registered on the platform are writing the reviews without using the product or their services. These types of reviews are also counted as fake reviews. Due to the rise in technology, several bots have now the abilities to write a fake review on the platform that can affect the purchasing pattern of the user. The reviews posted on the platform play an important role as customers decision are influenced by them.
The fake reviews can be categorised into three types:

- Reviews that are posted to intentionally damage a products or platform reputation or to promote the services. These types of reviews are tough to classify from the real reviews as they are similar to each other.
- Neutral review or advertisement that give no information regarding the review.
- Business owners who are generating reviews for their own product or services.

To analyse the reviews posted on the platform various machine learning algorithms have been used that can identify the fake and malicious reviews. The main detection algorithm that is used for this purpose was supervised learning algorithms but due to the lack of any reliable data, existing algorithms are only reliable upon the fake and non-fake labels for the model building. As the usage and dependency of online services in increasing, more data is generated on the daily basis. After researching, researchers felt the need to understand the emotions of the text generated is of importance as well. The transformers were incorporated as the solution to the problem. According to the research work several algorithms like SVM, Random Forest, XGBoost, Naïve Bayes have been used and have performed well. Due to the vast data generated on the day-to-day basis and the need to understand the emotions researchers started using LSTM, GRU to tackle this issue. In this report several machine learning algorithms like Logistic Regression, Decision Trees, Support Vector Machines, Naïve Bayes, Random Forest, XGBoost, Extra Trees with different feature extractions like CountVectorizer and TF-IDF Vectorizer will be used. Also, the best of deep learning models like LSTM, Bi-Directional LSTM, LSTM with attention layers, GRU, Bi-Directional GRU, GRU with attention layers are used.

1.1 Research Question

The main focus of this report is in comparison and evaluating the performance of different Machine Learning and Deep learning models.
RQ: *"To what extent can machine learning and deep learning based models be used to predict deceptive reviews?"*

1.2 Research Objectives

- Critically review the literature on sentiment analysis using machine learning and deep learning techniques.
- Implement and Evaluate the machine learning and deep learning models.

- Compare the results of the models.

2 Related Work

Textual Review analysis using the concepts of Sentiment analysis focuses on analysing and extracting insights from text data such as that found on social media platforms like Facebook and Twitter, as well as opinion-based platforms like Amazon's Amazon Opinions. It is crucial for businesses to have input into the development of their corporate strategy and a thorough understanding of how consumers view their products (Jagdale et al.; 2019). People's perceptions of the commercial entity of a firm can be gleaned by using computer algorithms to analyse their purchase behaviour. Some examples of representations for this item are people, events, blog posts, and product experiences. This page draws on customer reviews posted on Amazon.com and includes information on various types of cameras, laptops, mobile phones, tablets, TVs, and video surveillance systems (Jagdale et al.; 2019). All online goods companies now utilise sentiment analysis, which has recently exploded in popularity. As more people started using a product, manufacturers knew they had to improve it so that it could keep up with demand. The number of comments left by internet users who have used a website, blog, or shopped online has increased. Other shoppers thought about the reviews' comments while making their purchases. Using the principle of emotional analysis, businesses have worked out how to give precisely what customers desire. In other words, Sensory Analysis is a method of data analysis in which user reviews are processed and presented to the user after being analysed (Sadhasivam; 2019). Using machine learning, automated sentiment analysis can ascertain the tone of textual datasets. Amazon.com product reviews, which employ the same sentiment analysis methods as those on other sites, may be a great place to evaluate their efficacy (Guner et al.; 2019). In this research, the author and his colleagues introduce MARC, a massive database of Amazon reviews designed for multilingual text classification. The reviews were collected in six languages over the course of eight years (2015-2019) to create the corpus. The reviews were collected in a number of languages, not only English. Each record includes the review's full text, title, stars, anonymous reviewer ID, anonymous product ID, and coarse-grained product category (e.g., "books," "appliances," and so on). Twenty percent of reviews in each language have been given each of the five possible star ratings. In all, there are 200,000 English teaching, development, and test sets, 5,000 Spanish teaching, development, and test sets, and 5,000 French teaching, development, and test sets (Keung et al.; 2020). Two different machine learning methods will be examined to see which one best captures the tone of Amazon consumers' reviews. The final point is that shoppers may learn more about the quality of a product by reading user reviews. Product rankings will be positively impacted by several features of product reviews. Consider factors like product quality, content, review length in relation to product lifespan, and the age of positive customer reviews. Manual processes are inefficient and time-consuming because of the volume of work they need. Nowadays, artificial intelligence researchers agree that machine learning is the most effective method for training a neural network (Dey et al.; 2020).

2.1 Research using the Machine Learning models for the Review Analysis

Some of the researchers using the concept of the text analysis for the review analysis used machine learning to determine if a review was positive or unfavourable. The findings of this research suggest that machine learning approaches may improve the categorization of Product Reviews. Compared to the Support Vector Machine's 93.54% accuracy, the Naive Bayes Algorithm's was 98.17% accurate (Jagdale et al.; 2019). It was unclear why it would be necessary to use these algorithms even if they were perfect. As a result, an ensemble method has been used to increase the reliability of the assessments. An ensemble classification method pools the results of many classification algorithms into a single verdict by tallying the votes cast on each approach and taking the average. In this study, they use a combination of the Naive Bayes, Support Vector Machine, and Ensemble methods. To improve upon the precision of the present technique, they recommended substituting it with an Ensemble strategy. Once this calculation is made, depending on user reviews, a product is suggested (Sadhasivam; 2019). Reviews, blogs, forums, and social media are just some of the platforms where users may share their opinions. People's opinions may be heard in many places, from app stores to travel sites to product reviews on Amazon. The customer can either provide a number rating or provide free-form feedback on the goods. The effectiveness of each algorithm differs depending on the specifics of the situation. Depending on its efficiency, precision, and the quality of the data it was trained on, each algorithm has its own set of pros and cons (Sadhasivam; 2019). The suggested method proposes using an Ensemble methodology to categorise the literature. When referring to ensemble voting, the term "majority" is commonly used. The required result can be achieved by combining many algorithms. Predictor performance determines the outcome of each algorithm (mode values of all the algorithms). Collectively, Naive Bayes and Support Vector Machines form the Ensemble method. Because of this, the proposed method would provide more reliable outcomes than do currently used algorithms. By consolidating the essential methods into one, an ensemble technique increases precision. Then put it simply, this approach is superior to the alternatives. The ultimate output prediction is based on the sum of all the models' predictions. Since no one prediction receives more than 50% of the vote, the model may conclude that no ensemble prediction is reliable (Sadhasivam; 2019). The researchers perform thorough sampling, filtering, and text processing to the documents to limit the amount of noise in the dataset, and then they supply many samples for training models in six languages with well-defined training, development, and test splits. They uncovered several holes in previously available multilingual corpora, which are now filled thanks to this study. Preparing data for cross-lingual text classification requires knowledge of how to assess the distribution of key features in the corpus and how to organise the data (Keung et al.; 2020).

Learning from hundreds of comments would be greatly simplified with a framework for polarising these assessments and gaining insight. In the initial stage of this inquiry, consumer sentiment was evaluated using the Naive Bayes classifier. Human emotions have recently been classified using a support vector machine (SVM) into two groups (SVM). Before feeding the data into a network model, it was evaluated using pre-processing methods including term frequency (TF) and inverse document frequency (IDF). The purpose of this work is to identify a machine learning method that outperforms support vector machine (SVM) and naive bayes (NB) classifiers, which have been statistically evaluated

(Dey et al.; 2020). Currently, there are only a few of algorithms for aspect-level sentiment identification on certain domains that account for bipolar words (words whose polarity varies dependent on context) during analysis. This paper offers a novel method for assessing the sentiment of individual aspects of an item, as opposed to the sentiment of the thing. The work has been constructed and tested based on Amazon customer reviews (crawled data), with emphasis placed on selecting aspect phrases for each review before considering the remainder of the review. When a dataset enters the system, it goes through a series of pre-processing stages designed to eliminate any extraneous data before being given a good or negative evaluation. Stemming, tokenization, casing, and removal of stop words are all components of this process (Nandal et al.; 2020). This strategy should be able to handle the influx of reviews. Using five key supervised learning classifiers including NB, LR, SentiWordNet, RF, and KNN, the comments should be classified as positive, negative, or neutral. Besides reporting their findings, they also discuss the difficulties they encountered. This research demonstrates that modern algorithms make full use of feature extraction and sentiment analysis while processing large numbers of reviews on Amazon, using both the internet and mobile devices. Definitions, information extraction and retrieval, machine learning's function, and the mining of user comments were all discussed (Dadhich; 2022). This work uses a number of machine learning methods, including as support vector machines, naïve Bayes, logistic regression, decision trees, random forests, and stochastic gradient descent, to classify Amazon product reviews for electrical parts into positive and negative categories (SGD). Decision trees have a 73.3 percent accuracy in making predictions, but logistic regression achieves an impressive 83.89% (Urkude et al.; 2021). The likelihood of a subsequent occurrence may be calculated using Bayes theorems by comparing them to those of earlier events. Applications of NB in text classification and spam detection were indicated by (Daniel; 2022). They start with only one observation, then separate out the relevant aspects, and then place it into one of several categories. A generative classifier, NB assigns input to one of many predetermined categories. Since NB is the simplest method and just requires a small dataset for training, it can process data quickly. To make decisions and forecast the future, Bayes's theorem is applied (Daniel; 2022). There is a near-universal consensus that all the reviews uploaded to various websites are fake. Data Mining Classification is utilised to establish if a user review is a spam or not. Several available Text Classification Algorithms have been proven to benefit from hybridization of classifiers. The purpose of this research is to develop hybrid classifiers for identifying spam in review submissions. The first step of the Classification-Classification process employs base classifiers such as Naive Bayes and K Nearest Neighbor (KNN), while the second step employs a Support Vector Machine (SVM) classifier. Using data from both Amazon and Yelp reviews, the suggested Hybrid classifier's accuracy improves from 89.04 percent to 93.50 percent, leading to a noticeable performance boost (Krishnaveni; 2022). Looking at the rating and the date of purchase might help you determine if a review is legitimate. They also hope to aid users in determining if a post is spam by classifying customer reviews into false and not fake categories based on review-centric criteria. Among many other considerations, one might consider the review's star rating, the product being reviewed, and the review's reliability (Kotriwal; 2022).

2.2 Research using the Transformers models for the Review Analysis

By optimising a multilingual BERT model with reviews data, we present ground-breaking results for supervised text categorization and zero-shot cross-lingual transfer learning. Given the ordinal nature of the evaluations, we suggest utilising mean absolute error (MAE) instead of classification accuracy for this task (Keung et al.; 2020). Standard domain adaptation methods reduce differences between the source and destination domains to accomplish sentiment migration; however, they neglect efficient sources and can't handle a negative transfer, leading to subpar results. When developing a method for selecting domains from many sources, data quality can be enhanced by employing a contrastive transformer-based domain adaptation (CTDA) technique. The contrastive four-stage CTDA is presented as a method for constructing a discriminator to collect features' domain-private information through contrasting learning:

1. Creating a mixed selection that gives equal weight to all similar sources or only the Top-K sources that are equivalent in both domains in terms of space.

2. Extensive testing on two publicly available benchmarks shows that our CTDA model outperforms the state-of-the-art methods (Fu; 2022).

Sentiment analysis (SA) is among the many fields that focus on analysing and exposing insights from text data. It is critical for businesses to utilise this technology to assist them enhance their company goals and better grasp customer feedback on their products. The reviews in French used in this analysis were given by Amazon, which provided the data used in this analysis. We employed contextualised word embedding for features like ELMO, ULMFiT, and CamemBERT for French-language features before applying deep learning algorithms to classify reviews as positive or negative. The findings showed that the LSTM+CNN combination model trained using CamemBERT achieved a 93.7 percentage accuracy rate in classifying French reviews (Habbat et al.; 2021). These days, consumers may go online and read other people's reviews of products and services they're thinking about buying. A company's bottom line might be affected in the long run by these evaluations. Business reputations can be boosted or lowered by spam reviews, and the public might be misled in various ways. Online review integrity requires the detection and removal of fake reviews. Our phoney review detector made use of ALBERT, RoBERTa, and DistilBERT. Conventional machine learning and neural network-based models were overcome by our method. The performance of the models is evaluated based on their precision and their F1-weighted source weight. When it comes to identifying false reviews, the RoBERTa classifier provides superior performance compared to the gold standard model (Gupta et al.; 2021). Research shows that a BoW model is superior to utilising merely a Transformer for sequence classification, and that the classification errors of an imperfect BoW model shed light on the successes of a pure Transformer model. While our results aren't ground-breaking, they do demonstrate the need of using many models when doing data analysis. This is especially true when there is a scarcity of computer resources and the need to transform data into policy suggestions (Gupta et al.; 2021). As part of social media bot detection model, the researchers describe a novel approach to sentiment classification of tweets that makes use of Bidirectional Encoder Representations from Transformers (Google Bert) to unearth subject-independent characteristics. This study differs from earlier bot identification approaches in that it uses Natural Language Processing to build topic-independent characteristics for the new

bot detection model. Comparatively, the 82% accuracy of the best prior attempt was well behind the 94% achieved by Cresci et al.-2017-paradigms (Heidari et al.; 2022). Adjusting the BERT's parameters improves word representation, which in turn improves the reliability of emotional analysis classifications. To determine the likely orientation of a dataset, they employ a bidirectional Long Short-Term Memory classifier. To improve the efficiency of Bidirectional Long Short-Term Memory, APSO is used to choose appropriate weight values (Bidirectional LSTM). The effectiveness of the Bidirectional LSTM is boosted as a result. As a result of the enhanced self-attention mechanism included into BiLSTM, the user is free to zero in on the most pivotal sentences in any given situation. Four standard datasets were utilised in the trials for analysis of results (Shobana; n.d.). Misleading review-detection models were built on top of this framework using the bidirectional encoder representations from Transformers (BERT) method (Lee et al.; 2022). To that end, Google has published its pre-trained NLP model BERT in 2018. When it comes to gathering semantic content, BERT outperforms standard models that rely on static word vectors. BERT's ability to accurately identify polysemous words relies on the context in which they are used and makes use of dynamic feature vectors in the identification process. Unlike previous systems used for word segmentation, BERT can use Chinese characters as its basic unit, hence avoiding potential problems. The BERT approach is superior to other ways when working with Chinese text. It's also used to make up new datasets of false reviews for the model's training with the help of the textgenrnn model (Cao et al.; 2022). These are some of the most important things we hope to accomplish with our research: One, creating BERT-based models to detect fake Chinese reviews and evaluating how well they work in contrast to current approaches.

2.3 Research using the Recurrent Neural Network models for the Review Analysis

Using data from customer reviews posted on Amazon.com (N = 60,000), this research compares, trains, and analyses machine learning algorithms. Comparisons of accuracy were made between the MNB, LSVM, and LSTMM models (LSTM). When compared to other methods, the LSTM's accuracy and AUC were superior (both 0.90). An Amazon.com scraped data set of product evaluations from a variety of categories was used to evaluate the LSTM model's predictive abilities, together with 3.94 million reviews from Kaggle. Evaluations of furniture performed the best (accuracy = 0.92). Emotional content in product reviews may be categorised using LSTM networks, and these results hold true across different types of reviews. When there are more than two groups, further research is required to determine whether the classification is accurate (Guner et al.; 2019). Customer opinions are essential to businesses because they serve as a barometer of success. In addition, it helps consumers since it offers them an idea of what to anticipate from upcoming offerings. This research endeavours to analyse and compare several deep learning algorithms for accurately predicting user opinions on mobile phone ratings, as seen on Amazon.com and elsewhere. Projection values were derived from an analysis of these evaluations, which were categorised as favourable, negative, or neutral. Various methods have been developed and investigated, including long-term memory networks (LRNN), group long-term memory networks (GLRNN), recurrent unit gated unit (GRNN), and update unit (URU) (UGRNN). Glove, word2vec, and FastText by Skip-grams all employed word embedding as a feature extraction approach for sentiment analysis. Five algorithms using each of the three feature extraction methods are compared on a variety

of metrics, including accuracy, recall, precision, and F1-score, using both balanced and unbalanced datasets. GLRNN methods employing FastText feature extraction had the highest accuracy (93.75%) when applied to an imbalanced dataset. When compared to other methodologies, the published literature suggests that this discovery is the most precise. The LRNN algorithm achieved record accuracy of 88.39% on the balanced dataset (Alharbi et al.; 2021).

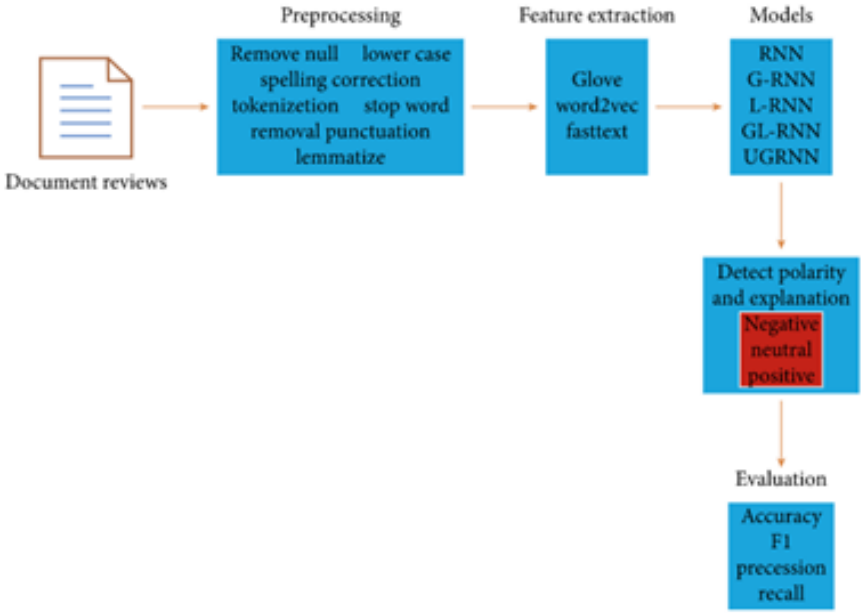

Figure 1: Process Flow used in (Alharbi et al.; 2021)

Some researchers employ a Bi-LSTM model equipped with an attention mechanism. One method they use to determine how easily something may be understood is by looking at how evenly emphasis is distributed throughout sentences and important words. In rare situations, the model's accuracy can approach 96%. It's interesting to see that aspect keywords get as much focus, if not more, than emotional elements in phrases (Li et al.; 2021). In this study, the researchers' look at how different types of machine learning may be applied to the Amazon reviews dataset. Bag-of-words, Tf-Idf, and glove were used to vectorize the reviews. Then, naïve bais, bidirectional long-short term memory, and bert. Following this, evaluations of the models' accuracy, f1-score, precision, recall, and cross-entropy loss function were conducted. Finally, we analysed the predictive accuracy of the top model. First, they did some multiclass labelling, and then we switched to binary labels (AlQahtani; 2021). Regardless, it's time-consuming to sift through all the feedback. Limitations in phrase length, shifts in textual sequence, and complex reasoning all contribute to the difficulty of accurately forecasting a product's mood. To overcome these obstacles, they developed a model using a Bi-LSTM Self Attention-based Convolutional Neural Network (BAC). To avoid future data sparsity issues, they employ pre-trained word embedding to reduce the dimensionality of the text representation. An attention mechanism collects n-gram features and prioritises those that are most relevant to the current situation. They also assign different values to words and phrases to zero in on the most crucial details. Classification features are trained using CNN and Bi-LSTM to collect semantic and contextual data. The overall performance of the BAC model is measured against a variety of standards. The F1 measure for the suggested model was

91% accurate, while the accuracy was 89% overall (Zhao et al.; 2021).

3 Methodology

In this section the description about the different methodology and concepts that can be used for the text classification for dataset will be given. Since the previous works have used different features extraction techniques like TFIDF along with the different machine learning models like SVM, Bernoulli Naïve Bayes, Multinomial Naive Bayes etc this research will also be using these techniques to have a detailed comparison. The advanced methods like Bi-Directional Stacked LSTM, Bi-Directional Staked GRU with attention layers will also be used for the comparison with the machine learning algorithms. KDD(Knowledge Discover in Databases) methodology us used in this research.

3.1 Dataset

The dataset that will be considered in this report are from Hotel Review Data (https://myleott.com/op- spam.html (Hotels Dataset)) and Kaggle Amazon Ratings Data (https://www.kaggle.com/datasets/bharadwaj6/kindle-reviews) where each reviewer has 5 reviews and each product has 5 reviews in this dataset. The data in this data set has been collected from may 1996 to July 2014. This data set is a small subset of data collection of the product reviews from Amazon Kindle category. For the Hotel Dataset, each of the 20 hotels in the Chicago region is reviewed, both positively and negatively. The data have been described in two investigations. In (Ott et al.; 2011), they focused on those that were positive, while in (Ott et al.; 2013), they addressed those that were negative. Both the dataset will be used to train the models and compare the results to give the detailed analysis on the same.

3.2 Data Pre-Processing

After the selection and collection of the data from their sources the next step is to perform data cleaning and data preprocessing. Data cleaning is required as the text data might have some noises, punctuations, emoticons or text in different cases. If the text preprocessing is missed and uncleaned data is fed into the model for analysis it will not be able to generate good results. The techniques that will be used in this report to clean the data are as follows:

1. Converting the reviews into lowercase.
2. Remove the punctuation as it do not hold any useful information for the analysis.
3. Removing regularly occurring words.
4. Spelling correction using Text Blob library.
5. Tokenization
6. Stemming
7. Replacing emoticons with their polarity.

3.3 Model Training

The next step after preprocessing is to split the dataset into test and train dataset. Sklearn can be used to split the dataset using the library method train_test_split(). The dataset will be split into 70:30 ratio. The classification task in this report implements various machine learning and deep learning algorithm after splitting the dataset. The training dataset is used to fit the model and the test data is used for evaluating the fit of the model.

3.4 Model Evaluation

After the model successfully trains on the training data and the models are assessed on the test data to check the fit, the evaluation metrics can be used to evaluate and compare the model. In this report the model will evaluated based on their accuracy. The model with the highest accuracy will be considered the best performing model.

4 Design Specification

In this section the overall design of the implementation process will be discussed, briefly describing the steps and tools used.

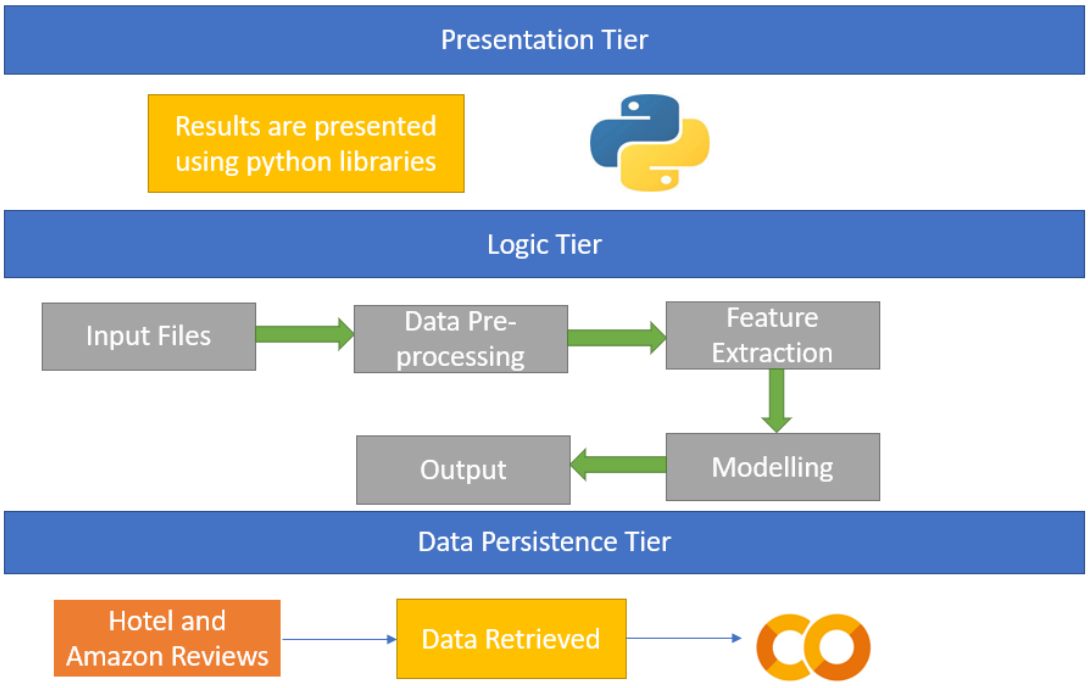

Figure 2: Architecture Design Used

In the data persistence stage, the data is collected from the sources. Both the dataset that are used in this project were available online. After the data collection, data was loaded into google colab for further operations and analysis.

In the logic tier,the procedure is followed in a sequence. First the data was loaded in google colab to carryout data cleaning and pre-processing such as removing punctuations, converting into lower case, spell correction etc. It is a necessary step before the data is

fed into a model as unprocessed data might not give good results. As the main of the project is to analyse sentiments and compare the models, the pre-processing step before applying classification algorithms was an important step. After the clean data is generated, features extraction techniques like Count Vectorizor and TF-IDF were used for the classification algorithms.

In this stage, the visualization of the outcome generated in the previous stage takes place. The visualizations are in the form of bar plot for the comparison, confusion matrix etc. All the visualization has been generated using python libraries in the google colab.

5 Implementation of Models

In this section the implementation, results, evaluation of the models that are used in the project are discussed. In this project various machine learning, deep learning algorithms are executed and the best performing model is selected on the basis of highest accuracy. Each model has been trained and tested on the same dataset. For the implementation of the algorithms various libraries like Numpy, Pandas, Sklearn, matplotlib, Keras, tenserflow have been used. Google colab has been used for the for the implementation as it provides free GPU which was required to train deep learning models.

5.1 Models Used for Analysis

1. Support Vector Machine
2. Random Forest Classifier
3. Logistic Regression
4. Adaboost Classifier
5. Decision Trees
6. Multinomial Naive Byes
7. Bernoulli Naive Bayes
8. LSTM
9. Bi-directional LSTM
10. GRU
11. Bi-directional GRU

5.2 Feature Extraction

For machine learning models to use the text data, feature extraction process is mandatory step to achieve better results as it increases the accuracy of the models by extracting features. In this research, TF-IDF and Count vectorizer are the two feature extractors the are used separately on each model to comapre the performance. The feature extractors transforms textual data into vectors before it is fed to the model. Unlike the Count Vectorizer that only focues on the frequency of the unique words in the text, TF-IDF

also provides the importance of the words. The overall comparison of both the feature extractors on machine learning algorithms can be observed in this report.

6 Evaluation and Results

To carry out the analysis, the report will be taking into the account two major tracks:

1. Inclusion of different models to check the analysis of the impact of different models towards the prediction of the Deceptive Reviews.

2. Inclusion of "different reviews" into the original Deceptive Reviews dataset to check the impact of different models.

In this part, different models like Logistic Regression, Naïve Bayes, Random Forest, Support Vector Regression, Decision Trees, AdaBoost, XGBoost, K-Nearest Neighbours with two important feature extraction techniques like Count Vectorizer and TF-IDF Vectorizers have been implemented. Apart from this other deep learning techniques like LSTM with Bi-Directional architecture, Stacked Architecture and with Attention Layers, GRU with Bi-Directional architecture, Stacked Architecture and with Attention Layers and Transformers like BERT is used for the detailed analysis.

6.1 Machine Learning Analysis with Count Vectorizer as the Feature Extraction

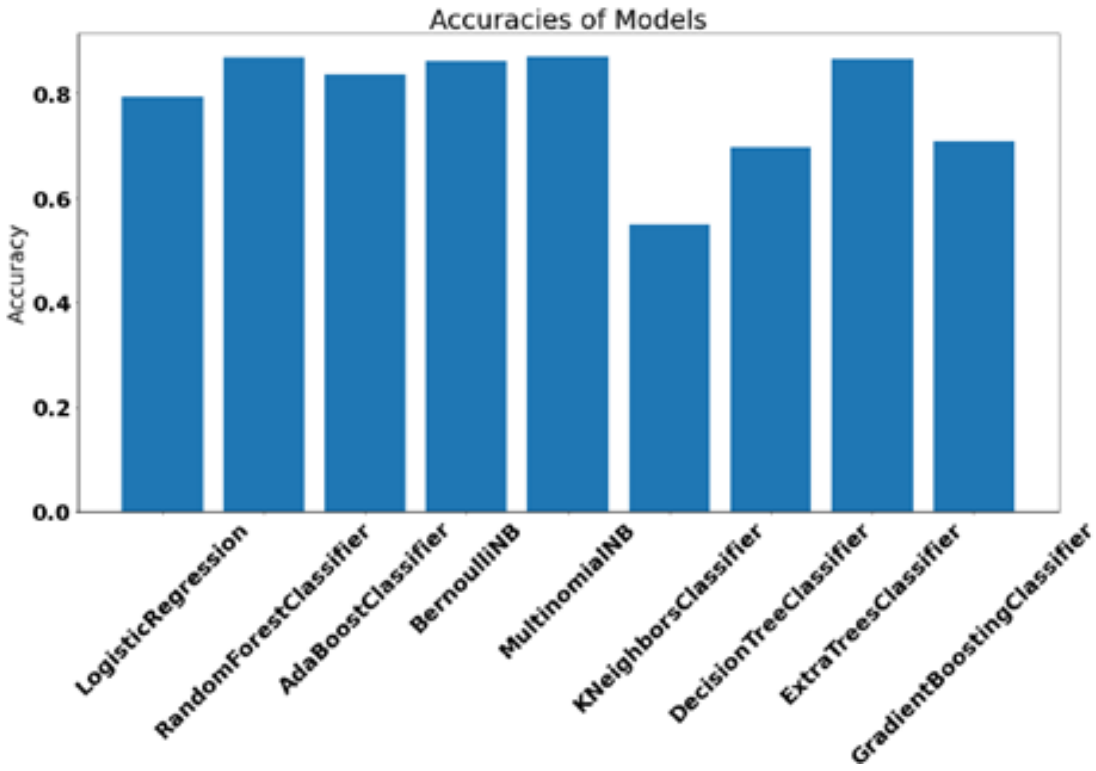

Figure 3: Comparison of Machine Learning Algorithms using

In the above graph, it can be found that the models like Random Forest, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Extra Trees performed equally well. The accuracy achieved by Random Forest is 86.87%, ¬Bernoulli Naïve Bayes is 86.25%, Multinomial Naïve bayes is 87.03% and Extra Trees is 86.67%. Hence Bernoulli Naïve Bayes performed the best in this bag of algorithms. To check the performance of SVM, the data that is fed into the system is binarized and then sent. Along with SVM, the other models that are used for this analysis is Bernoulli Naïve Bayes and Multinomial Naïve Bayes. The accuracies achieved by SVM is 80.45%, Multinomial Naïve Bayes is 88.21% and Bernoulli Naïve Bayes is 88.48%. Hence, from the usage of CountVectorizer, Naïve Bayes performed the best among all the algorithms.

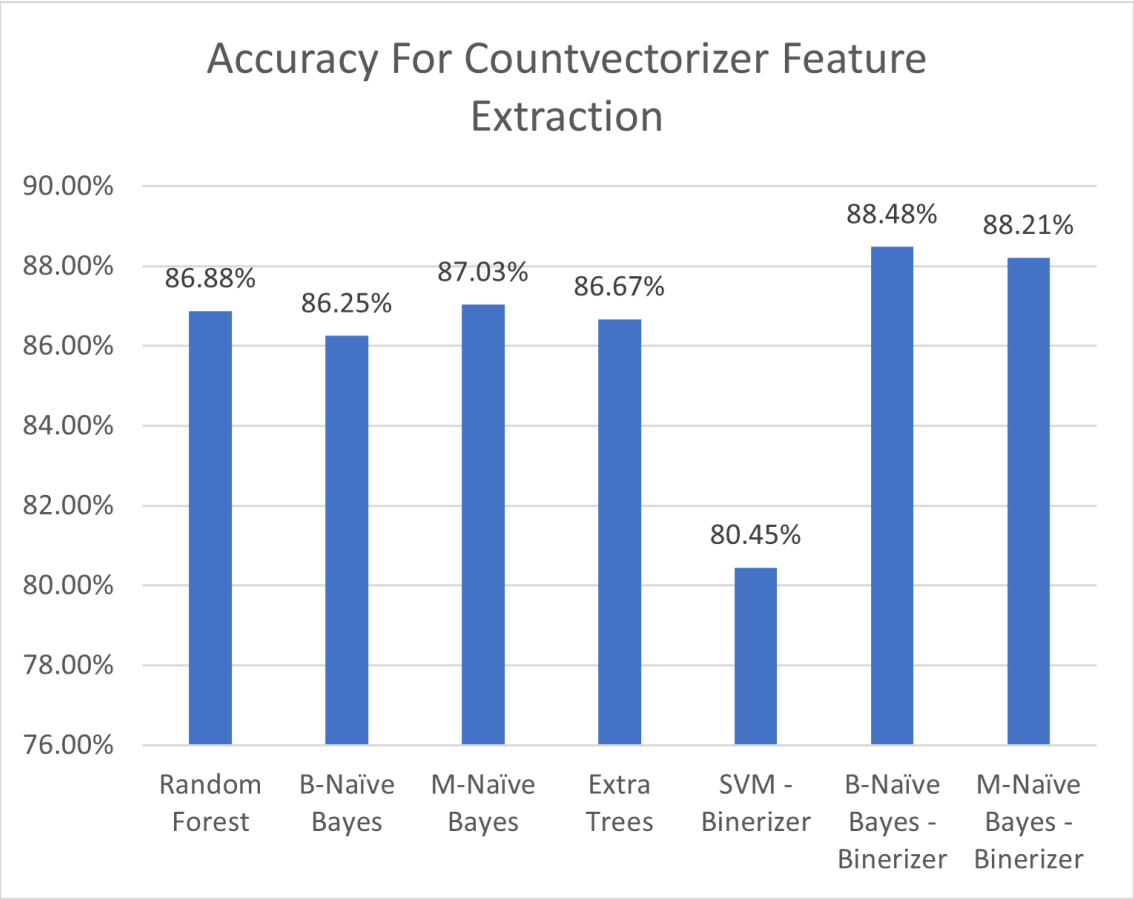

Figure 4: Analysis of the performance comparison using CountVectorizer

6.2 Machine Learning Analysis with TF-IDF Vectorizer as the Feature Extraction

Like the above analysis, when the TF-IDF is used, the following is the initial model comparison graph.

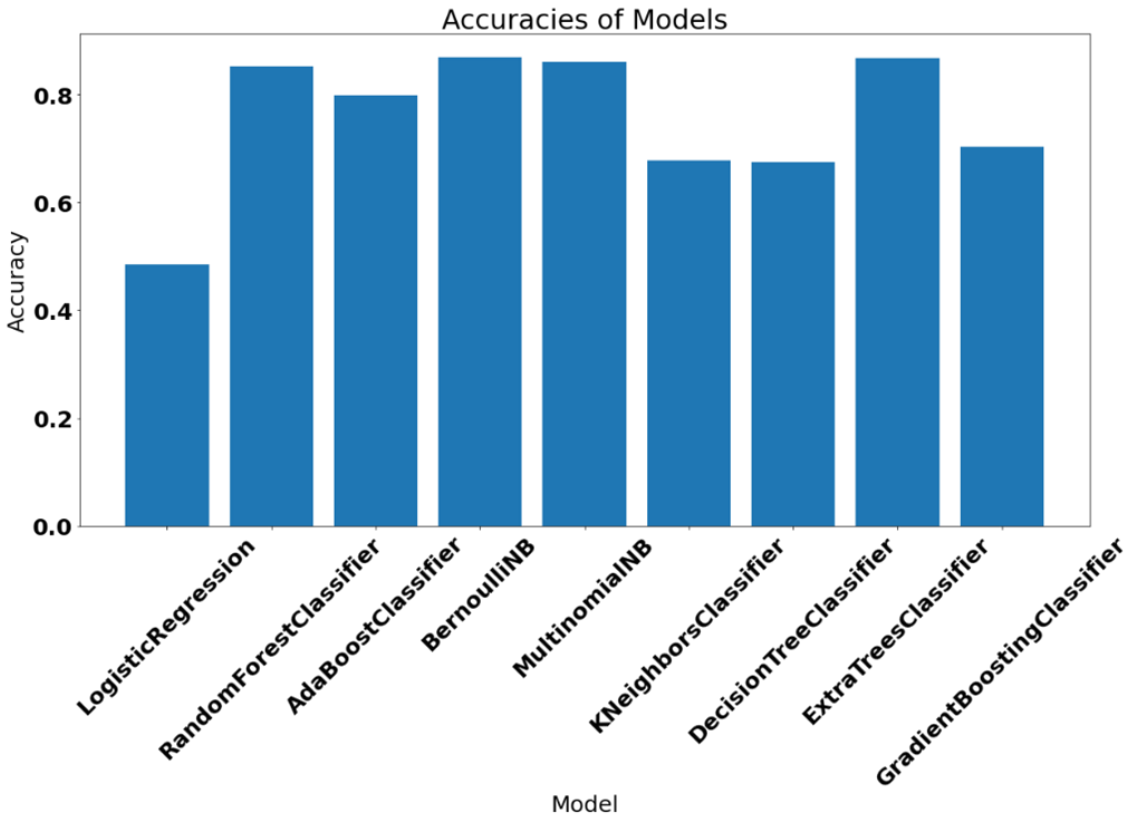

Figure 5: Comparison of Different Machine Learning Models using TF-IDF Vectorizer

As can be seen from the above graph, the algorithms that performed best are Random Forest, Bernoulli Naïve bayes, Multinomial Naïve bayes and Extra Trees. The accuracies achieved are 85.20% by Random Forest, 86.88% by Bernoulli Naïve bayes, 86.04% by Multinomial Naïve Bayes and 86.67% by Extra Trees. It is found that Bernoulli Naïve Bayes performed the best from this set of algorithms using the TF-IDF vectorizer. For checking the performance of SVM, the data is first converted to bag of words and then this is passed onto the TF-IDF vectorizer. In this analysis, SVM is found to achieve an accuracy of 86.31%. When Linear SVC is trained and checked the results, the accuracy achieved is 91.52%. A summary of the comparative results is as below,

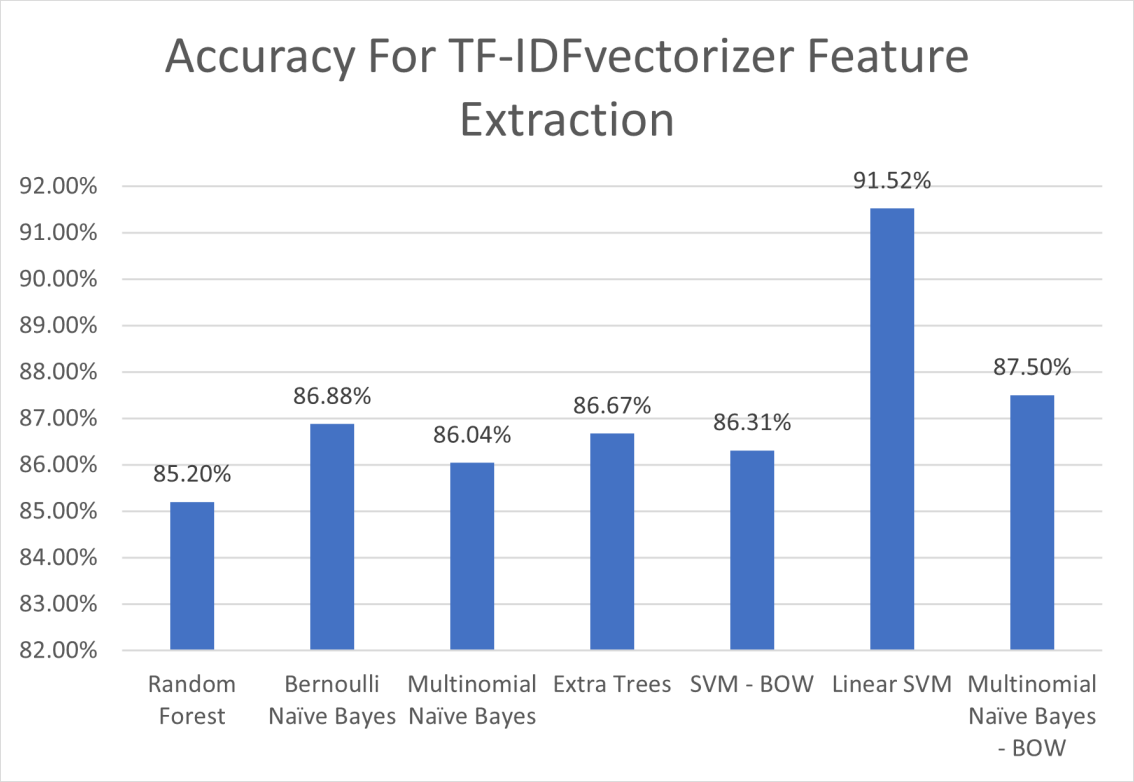

Figure 6: Detailed Analysis of the performance comparison using TF-IDF Vectorizer

From the overall analysis its found that Linear SVM using the TF-IDF vectorizer performs the best in the bag of models using the statistical features as the input vectors. The confusion matrix for the Linear SVM is as below,

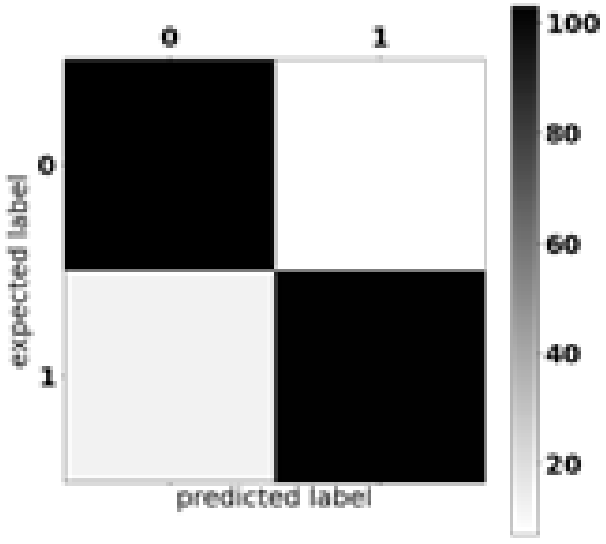

Figure 7: Confusion matrix for SVM

6.3 Deep Learning Analysis for the Deceptive Reviews Analysis

The batch size used for the analysis is 64 and the epochs used are either 20, 50 or 100. For the better learning of the models and to handle the overfitting and underfitting,

early stopping is used. The early stopping uses the Validation Loss to verify the stopping mechanism. The models used here are Bi-directional LSTM 2 layer with 64,32 Nodes, LSTM with 1 layer and with 50 Nodes, LSTM with Attention Layer. The summary of results are shown below,

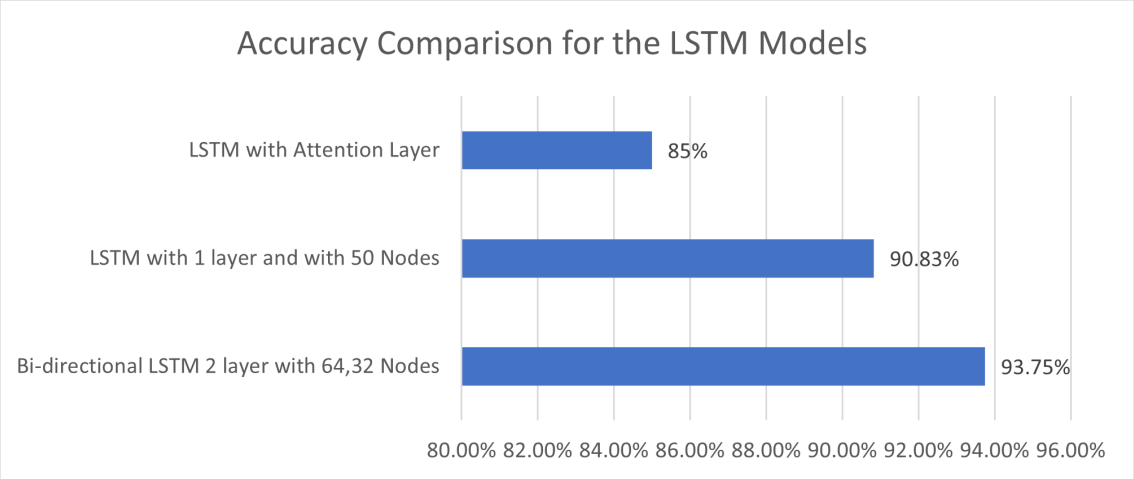

Figure 8: Accuracy Comparison of LSTM Models

Its is found that Bi-Directional LSTM is performing the best from the set of LSTM modules and the accuracy achieved is 93.75% at epoch 19. The history curve is as below,

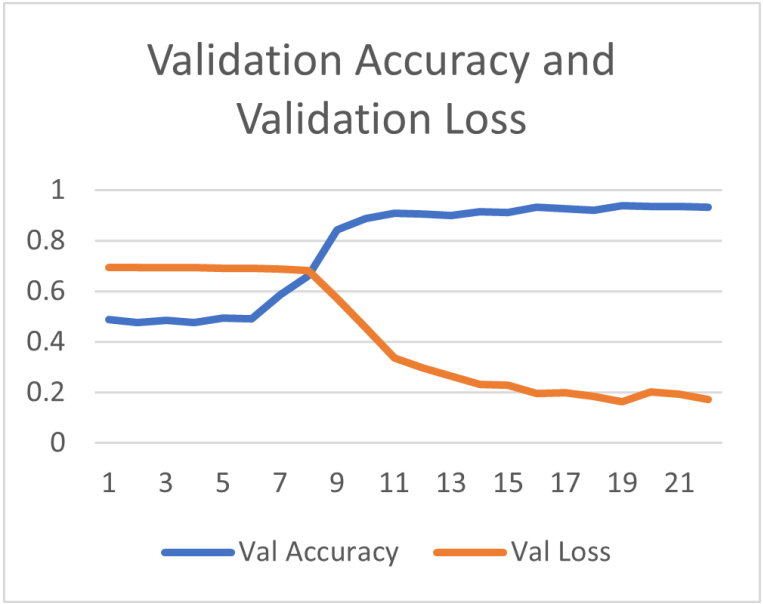

Figure 9: Validation Loss and Accuracy curves of LSTM and Bi-Directional LSTM

The GRU layers used here are Bi-directional GRU 2 layers with 64,32 Nodes, GRU with 1 layer and with 50 Nodes, GRU with Attention Layer. The summary for these models are shown below,

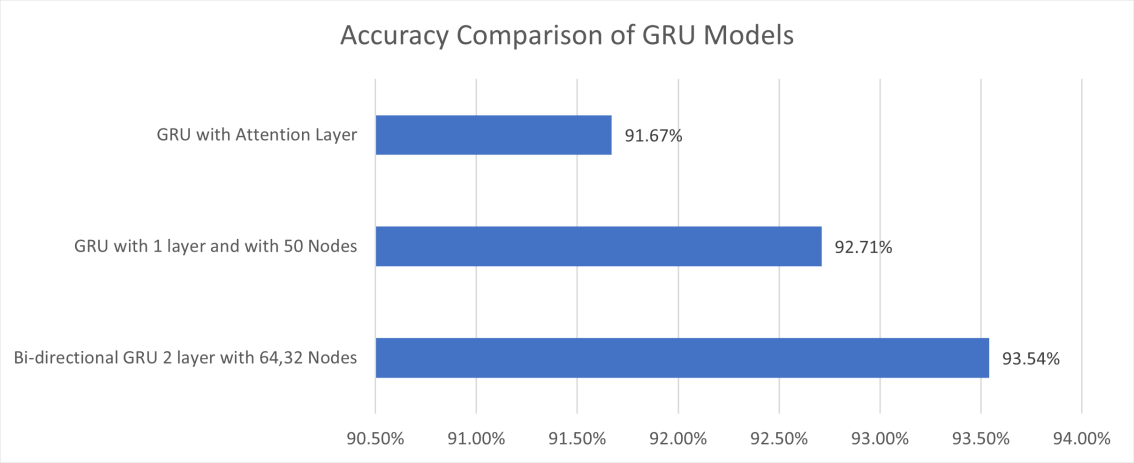

Figure 10: Accuracy Comparison of different GRU models

In this part the report follow the case 2 of the project.The analysis in this part will be done on the merged dataset.

Since the preliminary analysis is done, the best model is observed is the GRU bag of models. Hence, for the concept of inclusion of additional reviews from the dataset consisting of reviews from a different domain. The above data then merges the new amazon reviews dataset to the deceptive dataset reviews to make a superset of two dataset.

6.4 Machine Learning Analysis with Count Vectorizer as the Feature Extraction

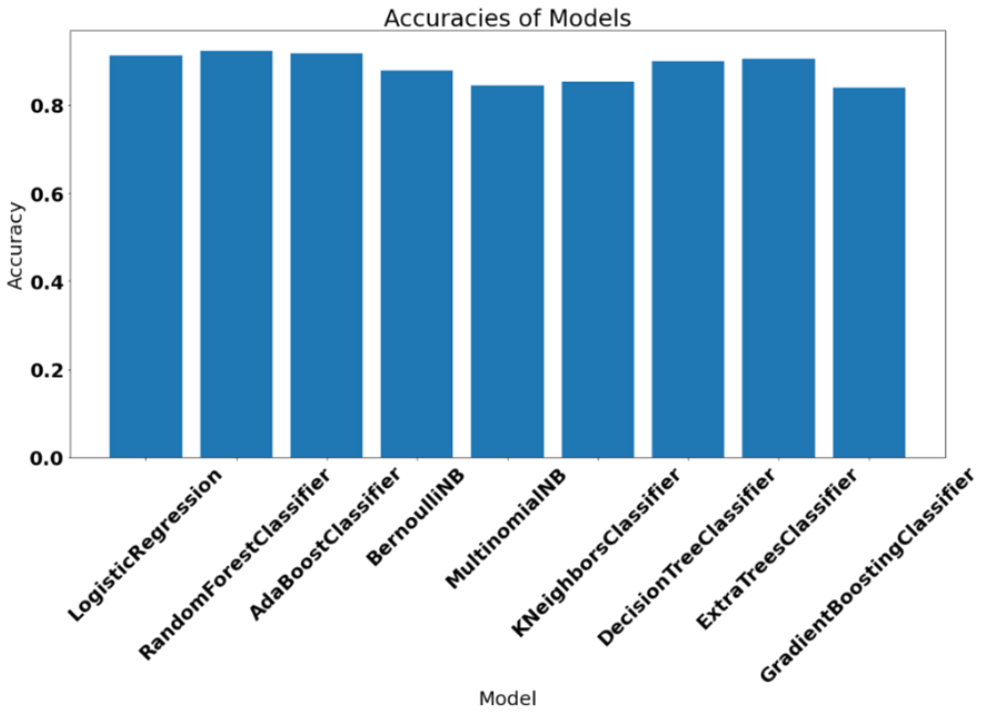

Figure 11: Comparison Analysis of Different Machine Learning Modules using the CountVectorizer

From the above graph using the Count Vectorizer, all the models are now performing well as comparison to the previous scenario. Logistic Regression got an accuracy of 91.30%, Random Forest with 92.27%, AdaBoost with 91.80%, Bernoulli Naïve Bayes with 87.80%, Multinomial Naïve Bayes with 84.5%, K-Nearest Neighbours with 85.34%, Decision Trees with 89.90%, and Extra trees with 90.52%. The reason is due to the inclusion of extra text that makes the models more training samples and more feature mapping set. SVM in this case didn't perform the best as in the previous case.

6.5 Machine Learning Analysis with TF-IDF Vectorizer as the Feature Extraction

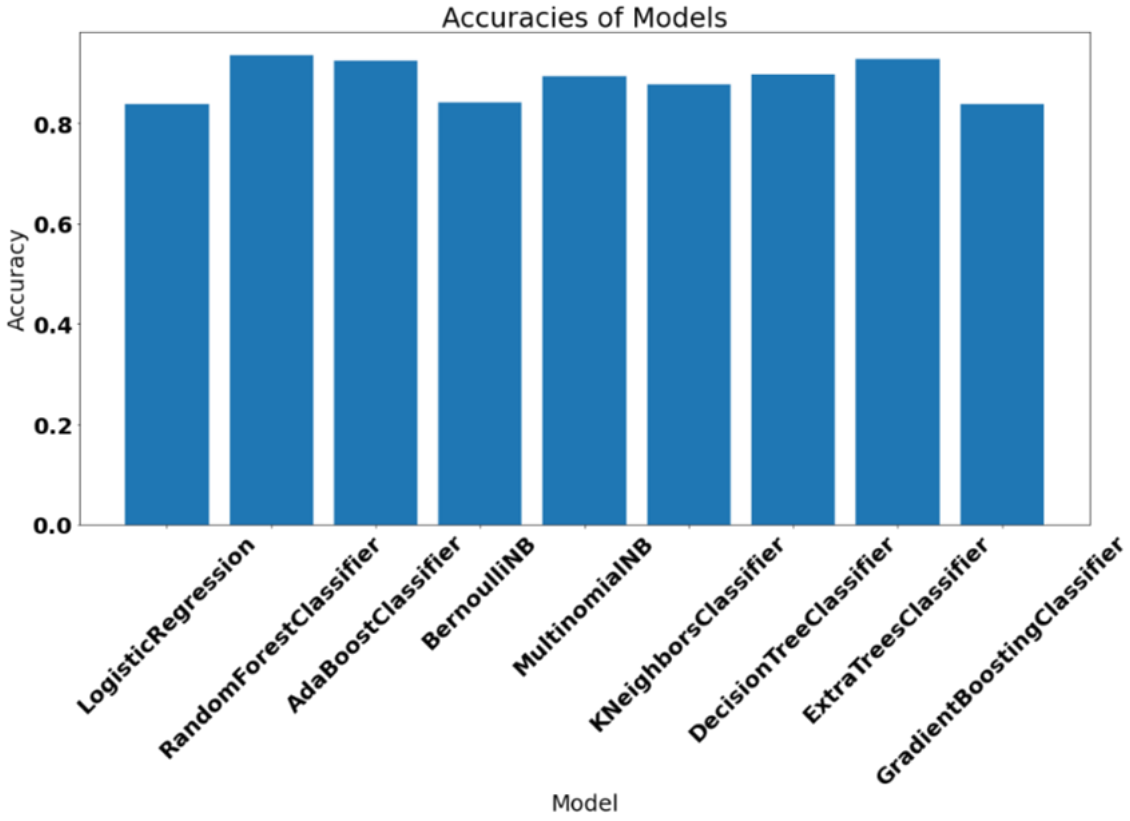

Figure 12: Comparison Analysis of Different Machine Learning Modules using the TF-IDF Vectorizer

In this case using the TF-IDF vectorizer, the performance by Random Forest is with 93.55% accuracy, Adaboost with 92.53% accuracy, Multinomial Naïve Bayes with 89.24%, K-Nearest Neighbours with 87.78%, Decision Trees with 89.78% and Extra Trees with 92.77% accuracy. In all the scenarios, Gradient Boosting didn't perform the best, but the performance improved with the inclusion of the new data. When SVM is taken into the account, TF-IDF with new data performs the best with an accuracy of 95.57% followed by Linear SV with 94.40% accuracy

6.6 Deep Learning Analysis for the Deceptive Reviews Analysis

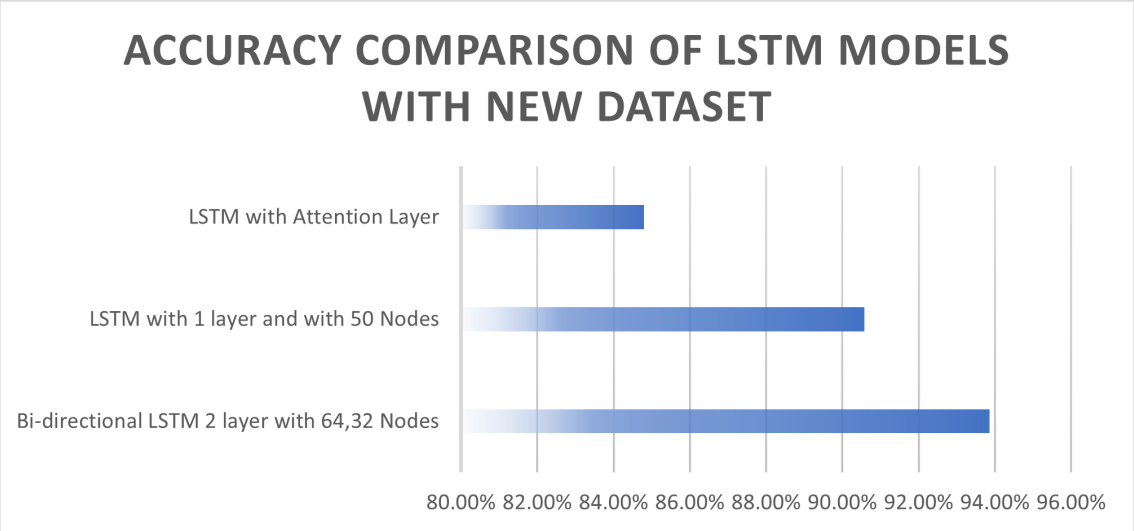

Figure 13: Accuracy comparison of LSTM models

Here, it can be found that Bi-Directional LSTM has improved the performance from 93.75% to 93.86%. Although the improvement is not very large, but the complexity is decreased with the number of epochs being used. The best performance is achieved with the epochs at 12. Hence the total time taken is also decreased.

In case of the GRU models, the performance is as below,

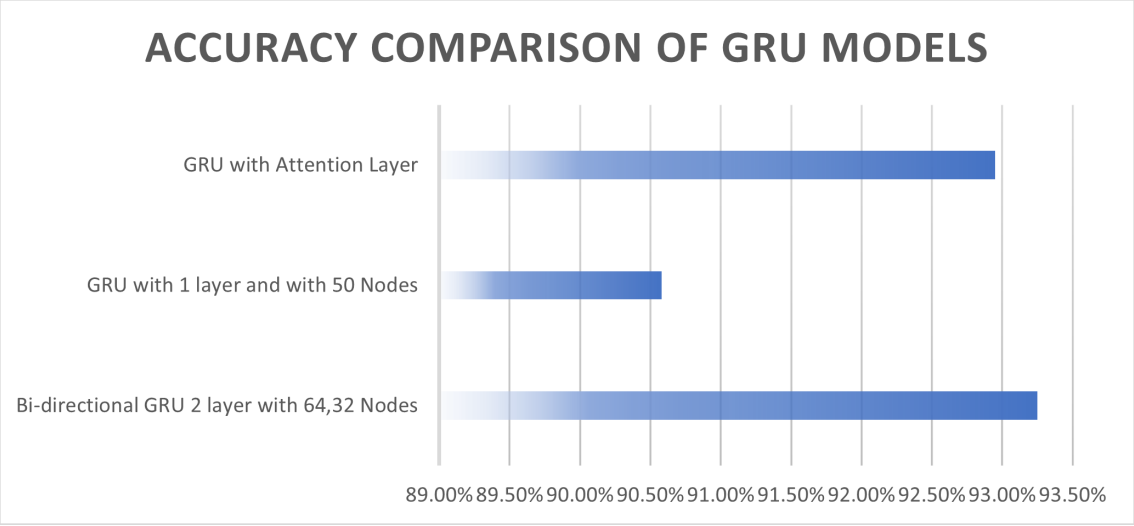

Figure 14: Accuracy comparison of GRU models

7 Discussion

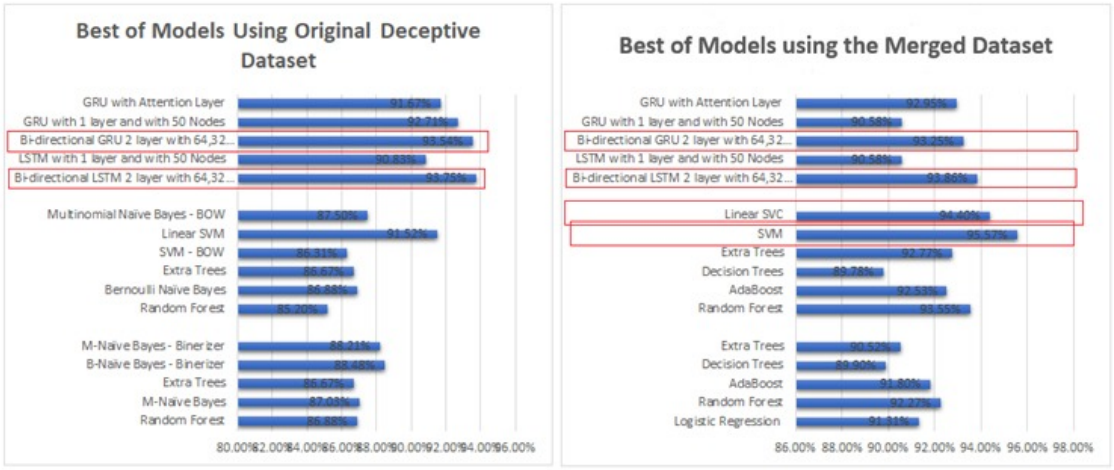

Figure 15: Comparison of Algorithms

After the implementation and analysis it can be found that the best of the models for the original dataset is Bi-Directional LSTM with 93.75% accuracy followed by Bi-Directional GRU with 93.54% accuracy. While the ReviewNet concept has achieved an accuracy of 95.57% accuracy with SVM followed by 94.40% accuracy with linear SVC. Bi-Directional LSTM and GRU have got 93.86% and 93.35% respectively. In other words, ReviewNet achieved an accuracy of 95.57% accuracy in comparison to 93.75% for the original dataset. The following points can be withdrawn from the analysis of the improving performance of the Case – 2 from the Case – 1. SVMs can generalise successfully in large dimensional feature spaces than the initial case having lesser feature space, which enables them to eliminate the demand for feature selection. A further advantage of using SVMs is that, in comparison to more conventional methods, they are more robust. Secondly, the TF-IDF method is a useful tool that analyses the frequency of words in a document to assess the importance of those words in the context of the document. It's a technique to weighing words that's not overly complicated but still makes sense. In the second case, the TF-IDF now contains more important information that makes the input vector for the SVM to work well and improve the performance. It can also be observed that in GRU has give the most stable output on the original dataset and as the data increases the accuracy of GRU decreases while the LSTM maintains it's accuracy as it is more complex and has three gates while GRU has only two gates.

8 Conclusion and Future Work

There is a widespread presence of false evaluations online in the context of online buying. The influence of online product reviews on consumer behaviour and sales has been the subject of a significant number of research studies. However, the existing body of research is mostly centred on genuine consumer product reviews, and just a small number of studies have looked at fraudulent reviews. This article analyses the elements that impact consumer buying choice in online review systems, which are swamped with fraudulent reviews. The results of the deceptive reviews were used as the basis for this article's

exploration. As a result, a model for the effect of misleading reviews is developed, and it is based on three key aspects of the online review system: the qualities of sentiment, the length of reviews, and the characteristics of online sellers. Text mining is utilised based on these to quantitatively analyse the indicators of the three key components.
It has been discovered that the model with the highest accuracy for the initial dataset is the Bi-Directional LSTM with 93.75%, followed by the model with the highest accuracy for the Bi-Directional GRU dataset with 93.54%. While the ReviewNet idea has attained an accuracy of 95.57% with SVM and then 94.40% with linear SVC, these numbers are not as impressive as they might be. The Bi-Directional LSTM and the GRU both received 93.35% and 93.86% of the total possible points. In other words, ReviewNet was able to attain an accuracy of 95.57%, but the original dataset only managed to achieve 93.75% accuracy. As a future scope of the work, the algorithms can be implemented on a larger dataset to oberserve the change in the accuracy. Also, different versions of BERT can be used on a larger dataset to further enhance the accuracy.

References

Alharbi, Alghamdi, Alkhammash and Amri, A. (2021). Evaluation of sentiment analysis via word embedding and rnn variants for amazon online reviews., *Mathematical Problems in Engineering* .

AlQahtani (2021). Product sentiment analysis for amazon reviews., *International Journal of Computer Science and Information Technology* **13**.

Cao, Ning, Shujuan, Dickson and Maoguo (2022). A deceptive reviews detection model: Separated training of multi-feature learning and classification., *Expert Systems with Applications* .

Dadhich, T. (2022). Sentiment analysis of amazon product reviews using hybrid rule-based approach., *Smart Systems: Innovations in Computing. Smart Innovation, Systems and Technologies* .

Daniel, J. H. (2022). Speech and language processing.

Dey, Wasif, Tonmoy, Sultana, Sarkar and Dey (2020). A comparative study of support vector machine and naive bayes classifier for sentiment analysis on amazon product reviews., *International Conference on Contemporary Computing and Applications (IC3A* pp. 217–220.

Fu, L. (2022). Contrastive transformer based domain adaptation for multi-source cross-domain sentiment classification.

Guner, Coyne and Smit (2019). Sentiment analysis for amazon. com reviews. big data in media technology., *KTH Royal Institute of Technology* .

Gupta, Gandhi and Chakravarthi (2021). Leveraging transfer learning techniques-bert, roberta, albert and distilbert for fake review detection., *Forum for Information Retrieval Evaluation* pp. 75–82.

Habbat, Anoun and Hassouni (2021). Lstm-cnn deep learning model for french online product reviews classification., *International Conference on Advanced Technologies for Humanity* pp. 228–240.

Heidari, Maryam and James (2022). Bert model for social media bot detection.

Jagdale, Shirsat and Deshmukh (2019). Sentiment analysis on product reviews using machine learning techniques., *Cognitive Informatics and Soft Computing* pp. 639–647.

Keung, Lu, Szarvas and Smith (2020). The multilingual amazon reviews corpus.

Kotriwal, S. (2022). Deceptive reviews detection in e-commerce websites using machine learning., *Data Engineering for Smart Systems* pp. 489–495.

Krishnaveni, R. (2022). A hybrid classifier for detection of online spam reviews., *Artificial Intelligence and Evolutionary Computations in Engineering Systems* pp. 329–339.

Lee, Yang and Lee (2022). Weight attention layer-based document classification incorporating information gain.

Li, Sun, Xu and Zhou (2021). Explainable sentence-level sentiment analysis for amazon product reviews., *5th International Conference on Imaging, Signal Processing and Communications* pp. 88–94.

Nandal, Tanwar and Pruthi (2020). Machine learning based aspect level sentiment analysis for amazon products., *Spatial Information Research* pp. 601–607.

Ott, M., Cardie, C. and Hancock, J. T. (2013). Negative deceptive opinion spam, *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pp. 497–501.

Ott, M., Choi, Y., Cardie, C. and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination, *arXiv preprint arXiv:1107.4557* .

Sadhasivam, K. (2019). Sentiment analysis of amazon products using ensemble machine learning algorithm., *International Journal of Mathematical, Engineering and Management Sciences* p. 508.

Shobana, M. (n.d.). An improved self attention mechanism based on optimized bert-bilstm model for accurate polarity prediction., *The Computer Journal* .

Urkude, Shubhangi, Vijaykumar, Urkude and C (2021). Comparative analysis on machine learning techniques: A case study on amazon product.

Zhao, Wang, Li, Liu, Yang and Liu (2021). Cross-domain sentiment classification via parameter transferring and attention sharing mechanism., *Information Sciences* pp. 281–296.