Classification of Hotel Reviews Using Sentiment Analysis and Machine Learning

by

Khalid Shifullah
18101062
Nuzhat Islam
18101374
Hasin Raihan
19301276
H.M. Rakibullah
18101371
Md. Ashik Iqbal
19341033

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
September 2022

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Khalid Shifullah
Khalid Shifullah
18101062

NUZHAT
Nuzhat Islam
18101374

Hasin Raihan
Hasin Raihan
19301276

H.M. Rakibullah
H.M. Rakibullah
18101371

Md. Ashik Iqbal
19341033

Approval

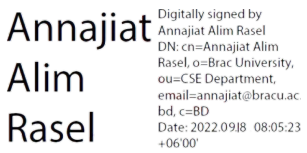
The thesis titled "Classification of Hotel Reviews Using Sentiment Analysis and Machine Learning" submitted by

1. Khalid Shifullah (18101062)
2. Nuzhat Islam (18101374)
3. Hasin Raihan (19301276)
4. H.M. Rakibullah (18101371)
5. Md. Ashik Iqbal (19341033)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on September 22, 2022.

Examining Committee:

Supervisor and Program Coordinator:



Annajiat Alim Rasel

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:

Dewan Ziaul Karim

Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:

Md. Golam Rabiul Alam,PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

We, herewith, state that this thesis is an original work of authors and has never been published before by anyone or any other institution. The results and analysis are achieved from our work. Co-authors and co-researchers are properly acknowledged in the paper for their extraordinary contribution. All used materials are cited carefully. We concur with the above statement and confirm that this submission complies with Brac University rules and regulations.

Abstract

Social media has become essential for people all over the world. It has given a platform for people to share thoughts, emotions, opinions, and ideas, causing a huge deal of data upsurge. Such an amount of data could be analyzed based on sentiment analysis and text classification via construction of an effective machine learning model. The concept gets more insight into it through analysis of the data, which is nearly impossible to conduct manually due to its huge configuration. This research focuses on the user's comments, and reviews about different hotels to predict their sentiment. As for the datasets, comments and reviews of hotels from online sites have been utilized. Moreover, text pre-processing techniques like tokenization, case folding, stopword removal, lemmatization, and duplicate data removal have been applied. TF-IDF and Bag of Words has been applied for word embedding. Furthermore, the effectiveness of supervised machine learning algorithms like, Support Vector Machine, Naïve Bayes, Random Forest, and Logistic Regression was evaluated and from the comparative analysis, it was observed that the Logistic Regression provided the most accuracy ranging from 86 to 89 percent.

Keywords: Sentiment Analysis, Word Embedding, Classifier, Tokenization, Decision Tree, Random Forest, Logistic Regression.

Acknowledgement

Initially, our appreciation goes to the honorable supervisor, Annajiat Alim Rasel, for providing all the guidance, advice, and encouraging us to initiate our research.
Secondly, we are grateful for the contribution of our Co-Supervisor, Dewan Ziaul Karim, for directing and helpling us in every stage of this journey patiently.
Our final gratitude goes to Brac University for this opportunity of conducting our study along with completing our undergraduate program.

Table of Contents

List of Figures

List of Tables

Nomenclature

The list below contains various symbols & abbreviation that will be used in the study

AI Artificial Intelligence

API Application Programming Interface

BoW Bag of Words

CART Classification And Regression Tree

CNN Convolutional Neural Network

Fig Figure

FNg False Negetive

FNt False Neutal

FP False Positive

IDF Inverse Document Frequency

ME Maximum Entropy

ML Machine Learning

NB Naive Bayes

NLP Natural Language Processing

NLTK Natural Language Toolkit

PSO Particle Swarm Optimization

RBF Radial Basis Function

RF Random Forest

SGD Stochastic Gradient Descent

SQL Structured Query Language

SSE Sum of Square Error

SVM Support Vector Machine

TF Term Frequency

TNg True Negetive

TNt True Neutal

TP True Positive

Chapter 1

Introduction

1.1 Background Information

At the time of booking a hotel, any potential client is often influenced by online reviews, ratings, and comments. Approximately 81% of people read reviews before booking a hotel [30]. Customer experience is crucial because it has an effect on a hotel's reputation and therefore the number of reservations it receives. Nowadays, online travel agencies and hotel booking services offer hotel ratings and reviews to customers. Consumer experiences can be analyzed by textual reviews or ratings given by customers. Among multiple factors that must be considered for a hotel to maintain a profitable revenue stream, its online presence and reputation are worth mentioning. Online reviews typically bear the most transparency as an indicator of a hotel's reputation. The hotel industry generally focuses on customer feedback to improve its operations, competitive positioning, and profitability. Reviews available on the internet have a more significant impact than hotel brochures on the customers. However, very few customers contribute actively with proper evaluations. Hotel booking services mostly publish scalar ratings, usually valued between 1 and 5. Customers give these ratings according to the fulfillment of their expectations.

Sentiment analysis implies the act of statistically distinguishing and classifying opinions listed in a text, particularly to establish whether the narrator has a favorable, unfavorable, or neutral viewpoint on a given subject. It is often called opinion mining, a branch of text mining and tongue processing research. Sentiment analysis helps data analysts in major organizations to estimate public sentiment, perform detailed market research, monitor brand and product reputation, and understand customer experiences. There are multiple machine learning algorithms to analyze the textual data like Support Vector Machine (SVM), Naïve Bayes classifier, MultinomialNB, Logistic Regression, Stochastic Gradient Descent (SGD), etc. Logistic Regression, Naïve Bayes Classifier, and Support Vector Machines (SVM) are examples of conventional machine learning approaches which are commonly used for sentiment analysis on a broad scale. The main focus of this research is to determine how customers feel about the hotel using sentiment analysis without having to read thousands of client comments at once. This is to eradicate the waste of manpower that might be involved in reading all the comments in case plenty of feedback is received. For such cases, sentiment analysis has been applied through the usage of natural language processing (NLP) to evaluate whether the data is classified under

either positive and negative segments or neutral ones. It is to mention that sentiment analysis has proved to be a great marketing tool that allows hotel managers to understand customer emotions in a better way and incorporate them into their marketing strategies. Customer loyalty, customer happiness, advertising, and marketing success are some of the critical aspects of product and brand awareness. The use of the internet and websites is rapidly rising to make millions of people inclined to social networking sites to exchange various types of information. Social media serves as a conduit for information from all around the world, which are mostly unsorted due to its unstructured format. The state of events, products, business, and politics is reflected in a variety of viewpoints. Opinions can be expressed directly, indirectly, or even implicitly. Moreover, there can be numerous unnecessary comments, words, symbols, etc. Here arises the challenge to express such data in a way that allows exact prediction when ultimate results for a purpose are expected, to which the usage of NLP provides a solution through a pattern for sorting. This study deals with reviews which are usually raw text with no specific structure. Collected data from Datafiniti and Tripadvisor have been analyzed by the algorithms. The accuracy of the predictions may vary for different algorithms, regardless of which, the goal is to analyze the efficiency of the classification models to achieve satisfactory accuracy of the predictions.

1.2 Research Problem

Businesses seek to provide the most necessary service based on consumer feedback in this contemporary day when everything is available only with a simple click. Running sentiment analysis on consumer feedback and reviews potentially makes it feasible. Only then, modern companies may be able to anticipate the best client needs and prepare marketing strategies for their products [5]. For a variety of reasons, people express their ideas online and on various social media platforms. Processing this data is typically done by hand. Even so, it will be expensive and will consume a great deal of manual energy and time. However, it is possible to achieve this technically, thanks to advances in technology and storage capacity. In this research, hotel reviews that reflect people's experiences in textual data for text categorization and sentiment analysis from people's viewpoints have been chosen. Natural Language Processing (NLP) and Machine Learning (ML), both branches of Artificial Intelligence (AI), are two of the most promising techniques to emerge in recent years. These techniques are capable of classifying text intelligently depending on the mood it conveys. Open-ended text can be automatically categorized into a variety of specified categories using the text classification method of machine learning. Text classifiers are capable of organizing, arranging, and categorizing almost any type of material, including text from the internet, articles, scientific research, and customer concerns [33].

In order for computers to analyze sentiment or feelings in the same manner that the human brain does, natural language processing may be difficult. Additionally, one should be aware of the diverse ways of expressing emotions in different languages [8]. The success rate of sentiment analysis is measured by how it provides better accuracy with human assessment. Human assessment is constantly evaluated using various criteria that consider accuracy and memorability for both the target cate-

gories of negative and positive messages. Several difficulties and problems should be addressed, including known challenges in sentiment analysis like negations, emoticons, and comparative tone problems. For instance, words with strong positive (+1) and negative (-1) polarity include love and hatred, even though word conjugations like "not so awful" can also indicate average. More than often, conjugations like this are left out, which could have significant alteration in the effectiveness of the model. An advanced Application Programming Interface (API) could be a solution to these challenges.

Furthermore, in the preprocessing and feature extraction section there are several steps to prepare the data for vectorization also known as word embedding [26] which are segmentation, tokenization, stopword, stemming, etc. Later, pre-processed data is vectorized which is the process to convert words to vector forms. To make this happen there are numerous existing techniques and some of the most popular ones are binary encoding, TF-IDF, word2vec embedding, Bag of Words, TF encoding, Latent semantic analysis encoding, etc. The Bag of Word model and TF-IDF have been used in this paper. For the next part, selected classifiers will be trained with vectorized data. From this point on, text classification can be approached in two ways, one is a supervised method and the other is an unsupervised method. Some popular supervised methods are Naïve Bayes, SVM, Maximum Entropy, Logistic Regression, and so on. All of these algorithms behave differently depending on the quantity or quality of data. It will be a challenge to find the proper combination of algorithms to form a model with satisfying performance.

1.3 Research Objectives

As a large number of individuals express their ideas and opinions on social network sites like Twitter and Facebook, several studies on sentiment analysis of social media data have been conducted. Sentiment analysis has several uses in industries, from business and marketing to tourism and technology. A machine learning based technique is less complicated than a knowledge-based approach since it does not require a predetermined database of all emotions. There is also a recommender system that will help us improve our model's user experience. Recommender systems are software programs that give viewers suggestions depending on several factors. These algorithms determine which product consumers are most likely to order and are most attracted to. The recommender system filters a large volume of data by focusing on the essential information based on the information provided by the user and other factors such as the user's desires and interests. To generate suggestions, it assesses the fitness of the individual and the item and the links between users and goods. It will help our research to give the best user experience. There are several machine learning algorithms and text classification techniques. This research paper will emphasize finding a model combining these algorithms and procedures to ensure maximum accuracy. This research is conducted to fulfill the following objectives:

1. To develop or optimize a model for sentiment analysis.

2. To evaluate the model.

3. To examine the relationship between sentiment, behavior, performance, and achievement.

4. To offer recommendations on improving the model.

Chapter 2

Literature

The expansion of the internet has given people chances to experience different products and services throughout the world, also to share feedback. Several websites were built to help customers to share their experiences with others. These feedbacks are essential for companies, brands, and other businesses to improve their services. However, the large number of websites having a big amount of customers make it difficult to go through the reviews manually, which leads business organizations to implement a sentiment analysis system. Emotional value can be detected from an enormous number of reviews with sentiment analysis classifiers. Various works can be found in this field to build models for detecting accurate emotion from users' shared reviews.

2.1 Sentiment Analysis

The word sentiment refers to a person's feelings regarding a particular object, fact, or experience. Sentiment analysis is detecting polarity by analyzing one's expressed feelings for something. The goal of a sentiment analysis system is to identify and classify emotions and opinions under some well-defined categories.

2.1.1 Approaches in Sentiment Analysis

The sentiment analysis process can be followed by two primary approaches [2]. The Lexicon-based approach perceives the sentiment of a document by detecting the semantic orientation of each word and phrase. Dictionary-based approach and corpus-based approach are the two subcategories of this approach. The corpus-based approach can be statistical or semantic. On the contrary, the machine learning approach is a "learn from experience" process that lets systems learn automatically without any human collaboration. The machine learning approach contains supervised learning and unsupervised learning. The supervised learning process trains classifiers with labeled text or sentences for future prediction [29], whereas unsupervised learning algorithms figure out the hidden structure of the given unlabeled data by itself [34]. Supervised learning can have four types of classifiers:

- Decision Tree classifiers: Random forest.
- Linear classifiers: Support Vector Machines, Neural Network.

- Rule-based classifiers.
- Probabilistic classifiers: Naïve Bayes, Bayesian Network, Maximum Entropy.

2.2 Related Work

This section talks about previous works on sentiment analysis which helps us to know more about this sector and to find the research gap, which is beneficial for further research.

Feature extraction is an important part of sentiment analysis, as it differentiates between the relevant and irrelevant features that the data contains. In research work [2] three feature extraction models and text preprocessing are presented. Text Preprocessing, which is a process that extracts features from the raw text by applying processes that include tokenization, punctuation, contractions, negations, lengthening, spelling, sentence segmentation, and stop-word removal. The bag of words model creates a vector representation of documents, where each term or word is a dimension in that vector space. However, the model cannot detect the order of terms in those documents. SentiWordNet is a sentimental dictionary model that gives relevant terms (especially adjectives) a score according to their strength of semantic orientation. For example, a negative score for negative words, positive for positive words. Then generates the average of the values. Word2Vec is another model that vectorizes words. The analysis used four classifiers, Support Vector Machines (SVM) with A Linear Kernel and an RBF Kernel, Maximum Entropy (ME), and Naïve Bayes with two different scales: 1-5 rating scale, and binary scale. The bag of words of the feature extraction model showed the best result where among the classifiers SVM model with RBF Kernel produced the best outcomes.

Many other researches show the outcomes of implementing different algorithms for sentiment analysis. The research work [20] was performed based on three different classification techniques: Naïve Bayes, Support Vector Machine (SVM), and Random Forest, with two different parameters to evaluate the performance. The feature extraction was followed by the TF-IDF model. The data was polarized into three different labels: positive, negative, and average. For Naïve Bayes, two different alpha parameters were used alpha=1 and alpha =0.009, Random forest classifier was implemented with 100 and 2000 decision trees, and lastly, the SVM was applied with class weight 1 and balanced weight. Naïve Bayes with alpha=1 could not predict the average label. But Naïve Bayes with alpha=0.009 performed with the highest accuracy. The result was evaluated using Performance Metrics.

Another research [18] compares five different machine learning models: Decision Tree, Naïve Bayesian, Logistic Regression, SVM, Neural Network, and Random Forest. The study was conducted over a dataset containing 10000 hotel reviews, among which 7800 reviews with 4-5 stars are labeled as 'positive'(1) and the rest as 'negative'(0). The dataset was prepared with a 10-folder-cross-validation design to avoid overfitting, where the dataset was shuffled and divided into ten sections. Nine sections of datasets were used for training and one for testing models. The first experiment included a few words, and the size of bytes has an average of 25.7

and a standard deviation of 14.6. Then, filtered reviews (eliminating punctuations and neutral words) were used for the second experiment, with a mean of 442.4 and a standard deviation of 394.0. In the earliest experiment, the Neural Network showed worse outcomes than the other five models. Other models gave 84%-87% prediction accuracy, among which SVM has the best result in recall, precision, and accuracy. For the second experiment, after using filtered data, the performance of the Neural Network model was increased by 8% but still worse than other models. On the other hand, SVM gave a maximum prediction accuracy of 92%.

Other researches show the outcomes of implementing different algorithms for sentiment analysis. Another study [22] shows the different results by executing four basic algorithms. Initially, their collected dataset was checked by Google Spell Check to correct misspelled words. Then, the texts were filtered by various data preprocessing tasks like removing punctuation marks, stop-words, unnecessary characters, calculating review length, stemming, segmentation, speech tagging, etc. However, this study observed four independent variables: quality of content, review sentiment, review recency, and reviewer characteristics. For the classification, Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes (NB) algorithms were applied. The dataset was randomly split into training and testing groups to evaluate performance. Accuracy, precision, recall, and F-measure scales were selected for evaluation as standards. In the case of outcome among 14,686 reviews, 10,158 were helpful, and 4,528 were unhelpful. Random Forest had the highest scores for accuracy, precision, recall, and F-measure among all used models, while Naïve Bayes (NB) had the lowest performance.

Convolutional Neural Network (CNN) is another approachable model used for text classification and image recognition [17]. This model consists of different layers, such as the convolutional layer, pooling layer, fully connected layer, etc., and uses SVM to perform the network processing. The outputs from three layers are connected to the Dense layer, where the ReLU activation function is applied and fed to a feed-forward neural network. Research shows using the CNN model over a dataset of 3000 reviews (2000: positive, 1000: negative) with 90 percent training set, and 10 percent testing set performs better than other models. The accuracy of detecting positive reviews was 95.345 percent, and for negative reviews, it was 96.145 percent. The overall accuracy was 98.22 percent.

Another research [12] presented a new system based on Factor Aggregation of sentiment polarization conducted in big data environments such as Apache Kafka as a data pipeline and for data processing, Apache Spark. The purpose of the constructed system is to combine hotel ratings and the sentiment degree of different hotel reviews. The study used hotel data and review data. Hotel data consists of 25 attributes, and review data has 12 attributes. As a language standardization approach, Yandex Translator API is used to translate reviews that are not in English. Collected review data are preprocessed by Tokenizing, Stop-word filtering, and Lemmatization. To determine the sentiment degree of hotel reviews, VaderSentiment and SentiwordNet tools are utilized, which use Lexicon-based and Rule-based methods. The SentiwordNet library uses a lexicon-based method, whereas the VaderLib library uses both methods. Mutations of both libraries are generated for the refined sentiment

degree. The outcomes of sentiment aggregation are passed as features for descriptive and predictive data analysis. For the descriptive analysis, the Hierarchical K-means clustering process is used, consisting of 8 features. This model acquired 4.61 percent of error performance using SSE (Sum of Square Error) in Sentiment aggregation. For SentiwordNet and Vader, the percentage is 6.12 percent and 7.25percent. In predictive analysis, the time-series model is used to predict the next month's value from the data. In this study, the baseline of 4 months with alpha = 0.8, gave the best outcome with an error of 12.4727percent. For other baselines, 5 and 6 months with the same alpha value, the errors are 12.5011 percent and 12.5316 percent.

The results of text classifiers are not often accurate enough. So, to optimize the output, there are various optimization methods to perform. The research work [16] has performed Particle Swarm Optimization (PSO) over the Naïve Bayes model with different parameters. The PSO model optimized the outcome of the Naïve Bayes model from 89 percent to around 91 percent. The optimal accuracy was achieved with 20-30 particles.

The above discussion shows multiple approaches to adopt for sentiment analysis. However, the previous research also shows some limitations. The research work [20] shows that data collection with the majority of positive data may create a bias in the result, as negative and average data is low. For data preprocessing and feature extraction, punctuation marks are removed. But punctuation marks may carry the emotion of the sentence [13]. So, by removing the punctuation, misleading comments might be created. Also, the customers may use native languages while writing reviews. The models may not capture the mixture of languages and leave the reviews ignored [22]. Besides, if a model is trained by datasets with a specific language, reviews that are written in other languages cannot be detected.

Chapter 3

Proposed Method

In this section, from data collection to analyzing the results will be discussed. Figure 3.1, is the workflow diagram which shows the techniques or models used in each section.

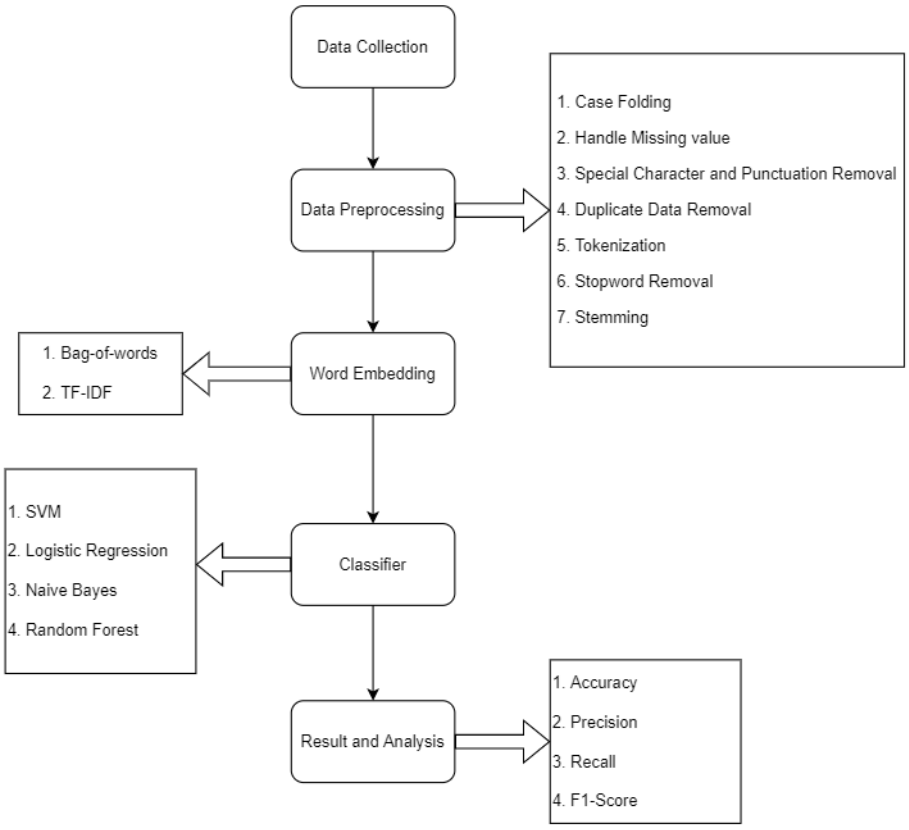


Figure 3.1: Workflow Diagram

3.1 Data Collection

Two different datasets have been used for this study. The first dataset is collected from Datafiniti [10] which contains hotel reviews of one thousand hotels. From this moment, this dataset will be addressed as Datafiniti dataset. The other dataset

contains hotel reviews from Tripadvisor website [6]. From this moment on, this dataset will be addressed as Tripadvisor dataset. The Datafiniti dataset holds a total of ten thousand reviews, whereas, the Tripadvisor dataset has a total 20,491 reviews.

3.2 Data Labeling

Both datasets contain review and correspondence ratings. The lowest rating can be given is 1 and the highest rating can be 5. Datafiniti dataset contains some rating in fraction value which is rounded by taking floor value. Table 3.1, shows the total number of reviews for each rating for both datasets. Figure 3.2 shows the class distribution of Datafiniti dataset. For the Datafiniti dataset, two labels will be used, positive sentiment and negative sentiment. Reviews with ratings of more than 3 are labeled as positive, and the rest of them will be labeled as negative. As for the Tripadvisor dataset, three labels are used which are positive, neutral, and negative. Reviews with ratings less than 3 are labeled as negative, reviews with a rating of 3 are labeled as neutral, and reviews with ratings more than 3 are labeled as positive. Table 3.2, shows the total number of reviews that are labeled positive, negative, or neutral for both datasets and figure 3.3 shows the class distribution of Tripadvisor dataset.

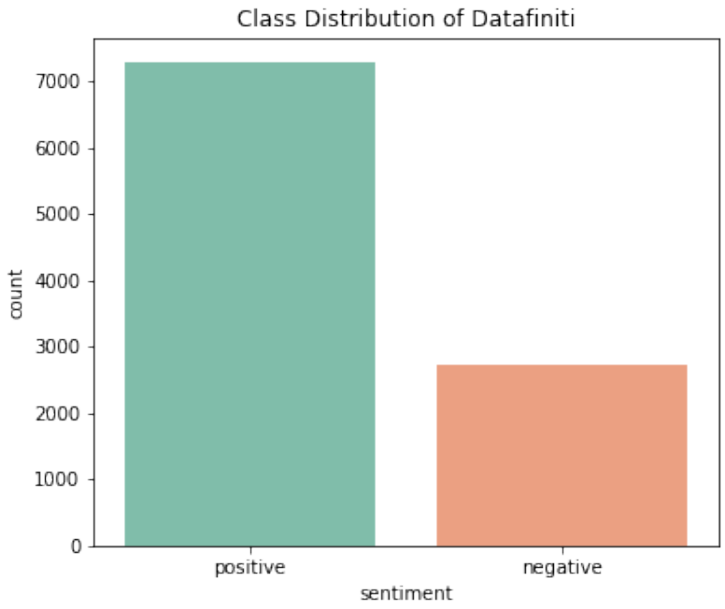


Figure 3.2: Class Distribution of Datafiniti

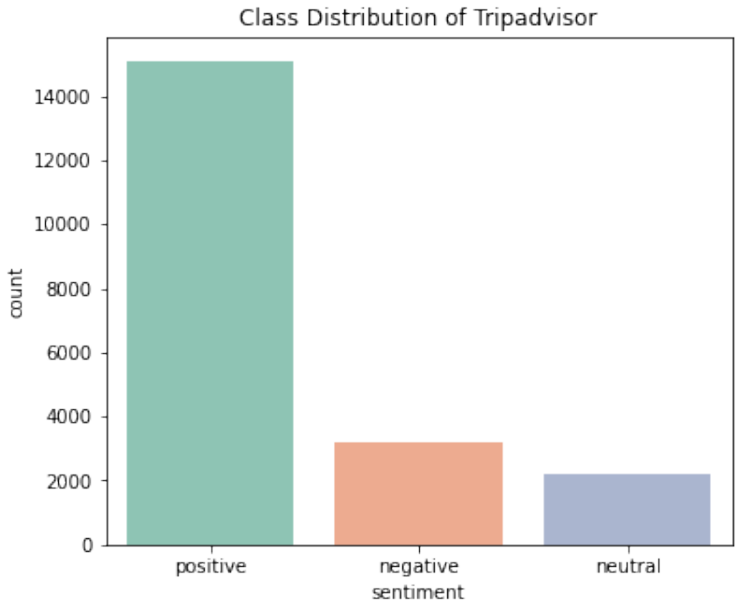


Figure 3.3: Class Distribution of Tripadvisor

Rating	Datafiniti Dataset	Tripadvisor Dataset
5	4,284	9,054
4	2,901	6,039
3	1,461	,2184
2	661	1,793
1	592	1,421

Table 3.1: Value Counts

Datafiniti Dataset		Tripadvisor Dataset		
Positive (>= 4)	Negative (<=3)	Positive (>=4)	Neutral (3)	Negative (<=2)
7,185	2,714	15,096	2,184	3,214

Table 3.2: Data Labeling

3.3 Data Selection

Long written narratives contain high sentiment. Short comments or comments with few words or single sentences do not express sentiment. Comments or reviews that contain less than 20 characters have been discarded.

3.4 Data Balancing

A balanced dataset has an equal number of positive and negative leveled data. In order to get higher accuracy for every algorithm, a balanced dataset or at least close

to a balanced dataset is very important. It ensures equal priority for every class. The Undersampling technique has been used to overcome data imbalancing. This technique balances the uneven data by keeping them in a minor class and reducing the actual size of the major class. This technique can be applied by randomly selecting data from the major class and discarding them.

3.5 Data Preprocessing

3.5.1 Case Folding

Case folding is a step of preprocessing where the texts are converted to lowercase. As a result, classifiers can deal with the same words with different configurations, like 'dinner' and 'Dinner' where both will be considered as a single word.

3.5.2 Handling Missing Value

Often datasets contain a lot of missing values which means values under a dataset can be found missing. There are many reasons behind missing values such as inaccurate SQL implementation, undefined values, data corruption, etc [15]. However, working with missing values is important as such datasets can decrease the quality of developed machine learning models. Most of the models work with the features of datasets and are unable to maintain missing values. Thus, to get proper output, it is essential to handle the missing values. There are different ways to handle missing values such as, deleting rows that contain several missing values, replacing a missing value with the mean of other values under that feature, using continuous values instead, create another method to generate outputs for missing values [24]. The size of the dataset is the main factor behind choosing the suitable handling process. Rows that contain missing values have been deleted.

3.5.3 Special Character and Punctuation Removal

Reviews of hotels contain different special characters like emoticons. These characters and punctuations mostly contain no sentiment for the models to work on. So, cleaning the dataset by getting rid of the special characters and punctuations is necessary. There are different ways to remove the punctuations and special characters. The most effective way is to use regular expressions.

3.5.4 Duplicate Data Removal

Data redundancy is an important problem in data cleaning. Duplicate data not only affects the storage, but also creates data pollution. Keeping the data integrity is also significant. Hence, rows that contain duplicate data are removed.

3.5.5 Tokenization

Tokenization is a part of natural language processing. Tokenization refers to breaking a text into multiple tokens. A token can be a word, symbol, or phrase, depending on the application. Tokenization is necessary as these tokens operate as inputs for

the other preprocessing steps. Natural Language Toolkit (NLTK) is a platform that helps to work with datasets from human language [14]. NLTK contains a few tokenizers. For instance: Regexp Tokenizer, TreebankWord tokenizer, and WordPunct tokenizer. Regexp tokenizer generates tokens based on the given regular expression as a parameter. For example, as an input text, "I can't work right now",if we set the regular expression parameter as ("[\w] + "), the output would be ["I", "can't", "work", "right", "now"]. For different parameters like ("[\w] + "), the collection of tokens will be ["I", "can", "t", "work", "right", "now"]. So, it works based on a given input expression. The TreebankWord tokenizer also develops tokens based on regular expressions but considers punctuations as tokens. For the previous text, the output of TreebankWord tokenizer is: ["I", "ca", "n't", "work", "right", "now", "."]. The WordPunct tokenizer splits the text as, ["I", "can", " ' ", "t", "work", "right", "now", "."]. These tokenizers are important because punctuation can hold different meanings, such as URLs, dates, email addresses, prices, etc. However, the mostly used tokenizer is the white space tokenizer, where tokens are separated by white space or newline characters.

3.5.6 Stopword Removal

There are some words that are very less important for analyzing the sentiment of a sentence. These words are very common and frequently used [23]. Stopwords can be labeled as low-level information. For instance, a few stopwords are a, an, at, the, and, is, has, and, etc. These words need to be filtered out to get a clean dataset as they do not carry sentiments. To have a cleaner dataset with better features, stopword removal is an important process. Natural Language Toolkit (NLTK) [1] library is used to remove stopwords from the text.

3.5.7 Stemming

There is always a common root (satisfy) for every inflected word (satisfactory) irrespective of the degree of the inflection. Stemming refers to a text normalization technique that is used to generate root forms of derived words [28]. In this process, the last few characters of a given string are removed. The stemming method has been used for data cleaning in this research work. In this experiment, the porter stemming algorithm has been applied for stemming. The words 'satisfying', 'satisfied', 'satisfy' and 'satisfy' carry the same sentiment but in a different form. Stemming will transform 'satisfying' into the base word 'satisfy' with small letters to help evaluate the word easier and better through machine learning.

Before Stemming	After Stemming
"enjoying", "enjoyed", "enjoys"	"enjoy"
"disgusting", "disgusted", "disgust"	"disgust"
"Likes", "liked"	"like"

Table 3.3: Example of Stemming

3.6 Word Embedding

3.6.1 Bag of Words

The first step in vectorizing a text is to build a dictionary of all the phrases it contains. The simplest way is to pick all the words in the document and convert each one to a vector space. One method for producing the document vector is to determine if a word is present in the document and then assign a boolean value to the relevant dimension of the vector. This approach is known as a "Bag of Words" because it works by verifying which words are in the bag after placing the words that make up a text [2]. There are terms that can be ignored since they really do not make a difference between documents (i.e. articles). In a corpus of hotel reviews, terms like "hotel", "service", "room" are probably used in the majority of the comments. Even if these terms are not likely to reflect specified sentiment towards the hotel that is the subject of the review, they might anyway be the subject of the sentiment analysis. The appearance of a word in a document can say a lot about a review. If a comment contains words with a direct emotional expression like "bad" or "good", there is a high possibility the emotion of the whole review goes along with the mentioned term's indication. But mentioning the same term (good or bad) does not make the review worse but it holds importance for the review.

3.6.2 TF-IDF

TF-IDF is the abbreviation of term frequency-inverse document frequency. It determines the importance of a word in a specific document. It retrieves the most essential part of the information. TF-IDF measures the number of occurrences of a particular word or token (t) in a single document (d) or corpus. TF generally determines the relative frequency of a token used in a document, and IDF determines how significant a given word (token) is. TF (term frequency) value and IDF (inverse document frequency) value must be calculated separately and used in a final equation to measure the final TF-IDF. The TF-IDF value increases with the frequency of a word or token (t) in a corpus, but is offset by the present data that contains the word in the document [4]. The TF-IDF separates the repeated terms by the number of terms in the corpus and weights each term in the document. Each word of the document has a unique TF-IDF value.

$$TF\ (t,\ d) = \frac{total\ number\ of\ occurrence\ of\ t\ in\ d}{total\ number\ of\ t\ in\ d};\ \ t = term,\ d = document$$

IDF (t, D) = $\log \frac{N}{1 + DF(d,t)}$; D = corpus of document, N = total number of documents

3.7 Train-Test Split

The datasets used in this study have been divided into 25 and 75 ratio where 75 percent is the training set, and 25 percent is the testing set.

3.8 Algorithms

3.8.1 Support Vector Machine (SVM)

SVM is one of the most significantly used supervised machine learning algorithms. It is used for classification, regression problems, anomaly detection, etc. This algorithm falls under linear classifiers. However, the primary use of SVM is classification [32]. There are multiple types of SVM methods such as SVR (Support Vector Regression) and SVC (Support Vector Classification) [19]. Finding a hyperplane that optimally separates the data points of distinct classes is the basic goal of the SVM classifier. By detecting the biggest difference between the two groups, higher accuracy can be achieved. The SVM classification technique produces an optimum hyperplane which is also called a decision boundary, that classifies fresh samples into various groups based on labeled training data (supervised learning). Then, using this line, additional data points are subjected to predictions. The maximum margin classifier, locates the line or hyperplane with the greatest distance from the nearest training data points among all classes. For example, a dataset can be used with two characteristics (x and y) and two classes as an illustration (0 and 1). This information may be seen by mapping it in a two-dimensional space and coloring each point in accordance with its assigned class. The illustration is given below:

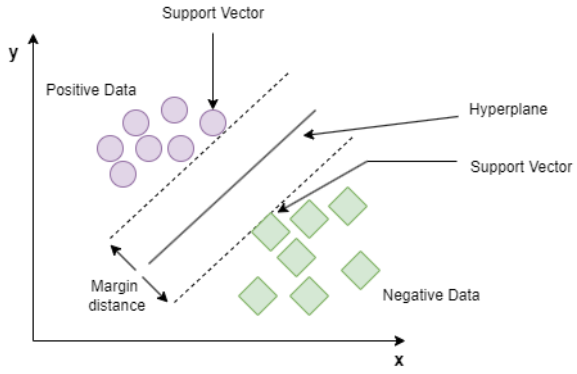


Figure 3.4: SVM Classification for Two Classes

In figure 3.4, positive and negative data are classified into two classes by the decision boundary. There is a distance shown for the nearest data point from both sets of data, which is called the margin. Greater marginal distance ensures better performance of the model. The lengthier the margin is, the less will be the error rate. There are several straight lines that can effectively divide the two classes. The SVM algorithm trains the model by locating the hyperplane (shown in figure 3.4) which divides the data into its two classes in the best possible way. Support vectors are the points nearest to this hyperplane.
Not all classification methods are linear. Real-world data is very complex. There often exists mixed data. These jumbled data are not linearly categorized. So, it is not possible to draw a single straight line to separate the distinct data.

For non-linear data, it needs to convert the 2D dataset into a 3D dataset. To distinguish a complex dataset, it has to be represented by a higher dimension dataset.

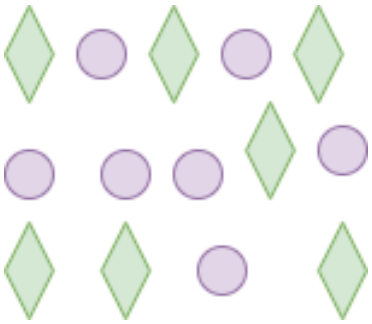


Figure 3.5: SVM Classification with Jumbled Data

SVM uses the technique called kernel function to convert the low dimension into a high dimension. Then, it can represent our dataset in 3D. But in this case, the hyperplane will not be a straight line anymore. To implement the SVM algorithm, the Scikit learn library is used. SVM model is the best for small-sized datasets as it needs a higher training time, also, works better with multi-featured datasets where feature number is more than data-points. To generate the decision function kernel functions are specified through customizing. However, this algorithm does not give probabilistic estimation but through five folding cross-validations which is costly.

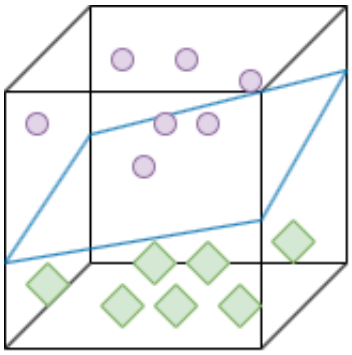


Figure 3.6: SVM Classification in 3D

3.8.2 Logistic Regression

Logistic Regression is a simple machine learning technique that belongs to the supervised learning category. It is a suitable technique when the output (dependent variable) is binary. It predicts the probability of categorical dependent variables. Not only that, but it is often preferred for its easy implementation process. It analyzes the relationship between categorical features (dependent variables) and independent variables by plotting them [27]. Instead of predicting a continuous variable like size, logistic regression predicts whether a given statement is True or False. Furthermore, logistic regression shapes the data to match an "S"-shaped logistic function rather than a line [3]. There is a logistic function called the "Sigmoid function". This function converts the independent variable into an expression of probability with respect to dependent variables. It measures values within the range of 0 and 1. The following is the equation of the sigmoid function.

$$y(z) = \frac{1}{1+e^{-z}}$$

z = independent variable
e = Euler's number= 2.718

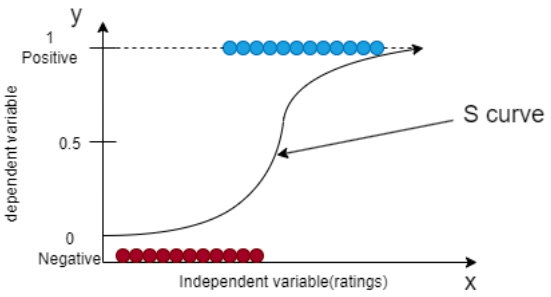


Figure 3.7: Logistic Regression-Sigmoid Function

The curve in the given figure 3.7 goes from 0 to 1, which means that it indicates the likelihood that a review is positive depending on the ratings. There is a good chance that the new review will be positive if a higher rating review is weighed. Finally, there is very little chance that a lower rating review is positive. Despite the fact that logistic regression may detect if a review is positive or not, it is most often applied for categorization. For instance, if the probability of a rating is greater than 0.5, then it will be classified as a positive review, otherwise, as "not positive".
Depending on the categorical response, there are three different logistic regression models. They are Binary logistic regression, Multinomial logistic regression, and Ordinal logistic regression.

3.8.3 Naïve Bayes

Naïve Bayes is a probabilistic classifier that uses Bayes algorithm [31]. This classifier accepts given features as independent of each other, which is often found false. In the real world, it is difficult to find a dataset with various attributes that are independent of each other [25]. But despite the false assumption, this classifier maintains to generate prediction models with a high accuracy rate. So, the classifier is often used to handle high-dimensional datasets. The Bayes law uses conditional probability that develops the chances of an event taking place on the condition of the occurrence of another event. It is defined as,

$$P(M|N) = P(M \cap N) \ / \ P(N)$$

Here,

- P(M|N) = probability of event M based on the occurrence of event N
- P(M∩N) = probability of the occurrence of event M and event N
- P(N) = probability of event N

Using this conditional probability Bayes theorem was stated :

$$P(M|N) = P(N \mid M)) \cdot P(M)/P(N)$$

- P(M|N) = Posterior probability that is the probability of event M after another event N is observed.
- P(N|M) = Likelihood probability which is the probability of N, given M is true.
- P(M) = Prior probability
- P(N) = Marginal probability

One of the simplest and fastest classification techniques is Naive Bayes. It is ideal for handling enormous amounts of data. It works well in many different applications, including:

- Recommender systems
- Sentiment analysis
- Text classification
- Spam filtering

For the prediction of unidentified classes, the Bayes theorem of probability is used. However, the disadvantage of this classifier is, due to the false assumption about feature independency, this classifier cannot understand the relation between features.

3.8.4 Random Forest

Random Forest classifier constructs decision trees. Multiple classification trees are included in it, and they may be utilized to estimate the class label. Every tree votes for a certain category label for a given piece of information and the category label with the most votes is added to the data point. In addition to the strength of a specific or individual tree within the forest, the error rate of this classifier depends on the correlation or relationship between any two trees within the forest. The trees should be strong and the level of associativity should be as low as possible in order to reduce the mistake rate. The internal nodes of the classifier tree are represented as the alternatives, the sides of a node are represented as tests on the weight of the feature, and the leaves are represented as category classes. Prior to detecting a leaf node, classification is performed starting at the foundation node and moving progressively below. The document is subsequently put into the class that corresponds to the leaf node.

Through a series of partitioning rules, the CART (Classification and Regression Tree) algorithm models the link between a response variable and a group of explanatory factors in order to reduce classification errors from each split [7]. All the data points would finally be classified as a consequence of several splits. The frequency of categorizing a data point properly (fi) or incorrectly (1-fi) may be computed for each split, and the degree of classification effectiveness overall would depend on lowering the Gini impurity. To determine Gini impurity, perform these steps:

$$I_G = \sum_n^i f_i(1 - f_i)$$

The construction of several trees in RF relying on a bootstrapped sample of the source information and the use of just a fraction of explanatory factors at each split are the two primary distinctions between CART and an RF method [9]. For most of the data points, roughly two-thirds are used in a single tree's bootstrapping for model construction. The sum of squares of the number of variables is employed at each split. It is ensured that the trees are all closely interrelated utilizing bootstrapping and a subset of the explanatory factors. In general, a single tree is more susceptible to the dataset's noise, which also leads to overfitting, subpar algorithmic efficiency, and subpar consistency when the data set evolves. When multiple uncorrelated trees are utilized in RF, noise and variance are reduced, improving algorithm performance.

Chapter 4

Result Analysis

In this study, accuracy, precision, recall and F1-score are used as metrics to measure the performance of the classifiers.

Accuracy:

Accuracy defines the percentage of correct predictions generated by the classifiers. It is the ratio of total true estimations and total generated predictions. Accuracy is the measure of effectiveness of a model.

$$\texttt{Accuracy} = \frac{TruePositive + TrueNegative}{TruePositive + FalseNegative + FalsePositive + TrueNegative}$$

Precision:

Precision is another performance measurement metrics which shows percentage of correct estimations of a predicted label. Higher precision ensures the perfection of the model's prediction.

$$\texttt{Precision} = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall:

Recall is the representation of the percentage of correctly predicted samples from a particular class. This metric also shows the effectiveness of the built model. The higher the recall is, the more capable a classifier is to predict the true sentiment from the samples.

$$\text{Recall} = \frac{TruePositive}{TruePositive + FalseNegative}$$

F1-Score:

F1-score is retrieved from recall and precision. F1-score is useful for the imbalanced class structure.

$$\text{F1-Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Table 4.1 and 4.2 represent the performance metrics for the models that have been used.

Classifier/ Word Embedding	Accuracy		Recall		Precision		F1-score	
	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF
Naive Bayes	77.9	76.7	77.9	76.7	77.45	95.58	74.19	56.32
SVM	87.55	88.9	87.55	88.9	87.79	89.75	83.99	85.27
Logistic Regression	**89.15**	86.85	**89.15**	86.85	**89.91**	89.13	**85.65**	81.78
Random Forest	83.65	83.85	83.65	83.85	88.75	89.54	75.88	75.85

Table 4.1: Performance Metrics (Datafiniti)

Classifier/ Word Embedding	Accuracy		Recall		Precision		F1-score	
	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF
Naive Bayes	73.23	73.31	73.23	73.31	70.68	70.65	57.91	58.31
SVM	83.55	89.24	83.55	89.24	85.23	92.83	68.47	76.07
Logistic Regression	**89.77**	87.75	**89.77**	87.75	**90.46**	92.69	**80.22**	72.28
Random Forest	83.23	80.99	83.23	80.99	94.41	94.95	59.11	52.18

Table 4.2: Performance Metrics (Tripadvisor)

4.1 Naïve Bayes

4.1.1 Naïve Bayes with Bag of Words

For both datasets, Naïve Bayes classifier provided decent results with the Bag of Words vectorization technique. Datafiniti dataset produced around 77 percent accuracy with 74 percent F1-score, whereas, Tripadvisor dataset generated around 73 percent accuracy with 57 percent F1-score. With smaller Datafiniti dataset, this model performed better.

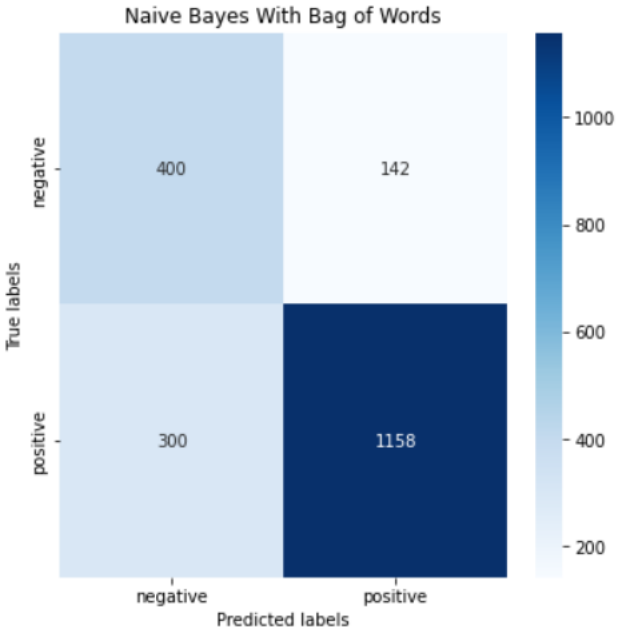


Figure 4.1: Confusion Matrix (Naïve Bayes + Bag of Words)

Figure 4.1 shows the confusion matrix of Naïve Bayes classifier with Bag of Words model for Datafiniti dataset. Total actual positive label counted from the figure

is 1458. Of all the true positive labels, 1158 of them are true positive labels and 300 are false negative labels. The number of total actual negative labels is 542 containing 400 labels that are predicted true negative and 142 are counted as false positive. Figure 4.2 represents Naïve Bayes confusion matrix with Bag of Words on Tripadvisor dataset. The total number of actual positive labels is 2988 of which 2443 are true positive (TP) labels, 260 label found as false neutral1 (FNt1) labels, and 89 are predicted false negative1 labels (FNg1). Total actual negative label is 640 where 408 labels are true negative (TNg), 134 of labels are false neutral2 (FNt2) and 98 labels are counted as false positive1 (FP1). The total number of actual neutral is 471, among that 103 label is predicted as false negative2 (FNg2), 151 are true neutral (TNt) and 217 labels are found as false positive2 (FP2).

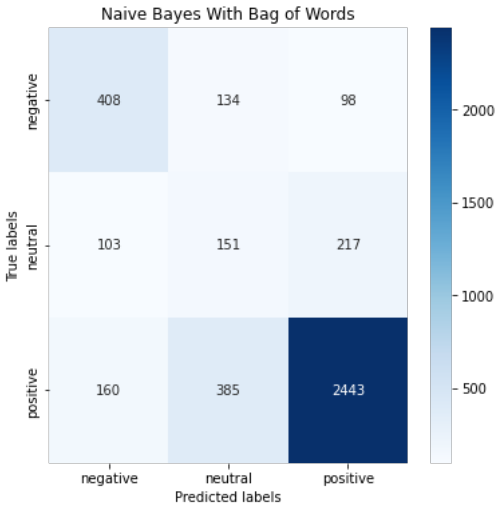


Figure 4.2: Confusion Matrix (Naïve Bayes + Bag of Words)

4.1.2 Naïve Bayes with TF-IDF

Naïve Bayes classifier underperformed with TF-IDF vectorization technique. Datafiniti dataset produced around 76 percent accuracy with F1-score of 56 percent, whereas Tripadvisor dataset generated around 73 percent accuracy with 57 percent F1-score. The Naïve Bayes with Bag of Words model performed better than this model. Figure 4.3 signifies the confusion matrix of Naïve Bayes classifier with TF-IDF over Datafiniti dataset. Total actual positive label is 1458 and 1450 of them are true positive labels and 8 are false negative labels. The total actual negative is 542 which consists of 84 true negative and 458 labels as false positive. Figure 4.4 represents the confusion matrix for Naïve Bayes with TF-IDF on Tripadvisor dataset. Total number of actual positive label is 2988 from which 2437 are true positive (TP) labels, 393 label found as false neutral1 (FNt1) labels, and 158 are predicted false negative1 labels (FNg1). Total actual negative label is 640 where 410 labels are true negative (TNg), 136 of labels are false neutral2 (FNt2) and 94 labels are counted as false positive1 (FP1). The total number of actual neutral is 471, among that 102 label is predicted as false negative2 (FNg2), 158 are true neutral (TNt) and 211 labels are found as false positive2 (FP2).

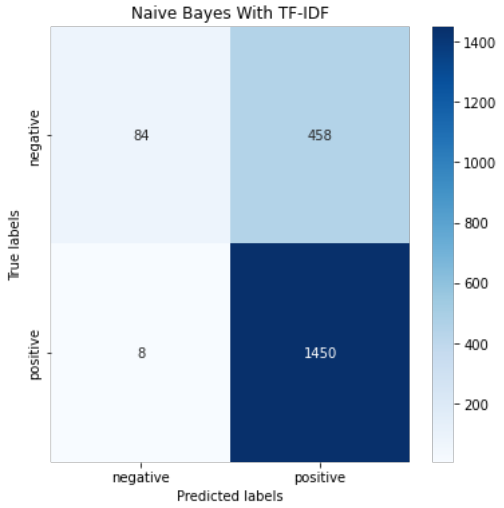


Figure 4.3: Confusion Matrix (Naïve Bayes + TF-IDF)

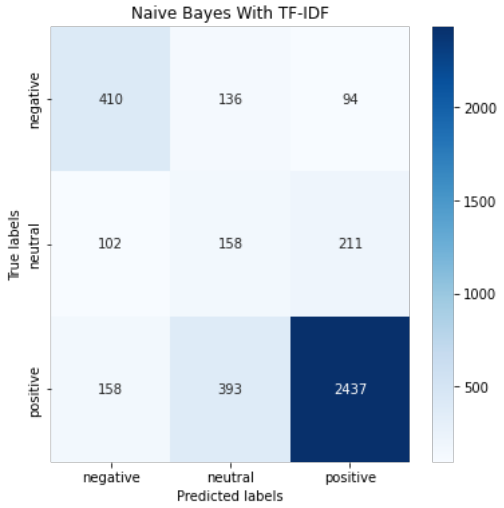


Figure 4.4: Confusion Matrix (Naïve Bayes + TF-IDF)

4.2 Random Forest

4.2.1 Random Forest with Bag of Words

Random Forest classifier provided satisfactory results with the Bag of Words vectorization technique. Datafiniti dataset produced around 83.65 percent accuracy with 76 percent F1-score, whereas, the Tripadvisor dataset generated around 83 percent accuracy with 59 percent F1-score. Again with smaller Datafiniti dataset, the model performed better.

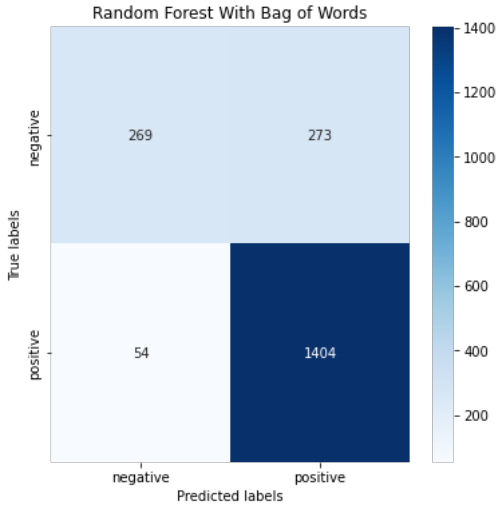


Figure 4.5: Confusion Matrix (Random Forest + Bag of Words)

Figure 4.5 shows the confusion matrix of Random Forest with Bag of Words model for Datafiniti dataset. The figure represents true labels and predicted labels. The number of total actual positive label is 1458 of which 1404 are true positive labels and 54 are false negative labels. Total actual negative label is 542 where 269 labels are true negative and 273 are counted as false positive. Figure 4.6 is a representation of the confusion matrix for Random Forest with Bag of Words model on Tripadvisor dataset. The table shows true labels and predicted labels. Total number of actual positive label is 3655 from which 3009 are true positive (TP) labels, 363 are false neutral1 (FNt1) labels, and 283 are false negative1 labels (FNg1). Total actual negative label is 399 where 359 labels are true negative (TNg), 30 labels are false neutral2 (FNt2) and 10 labels are counted as false positive1 (FP1). The total number of actual neutral is 45, among that 1 label is predicted as false negative2 (FNg2), 44 are true neutral (TNt) and no labels are found as false positive2 (FP2).

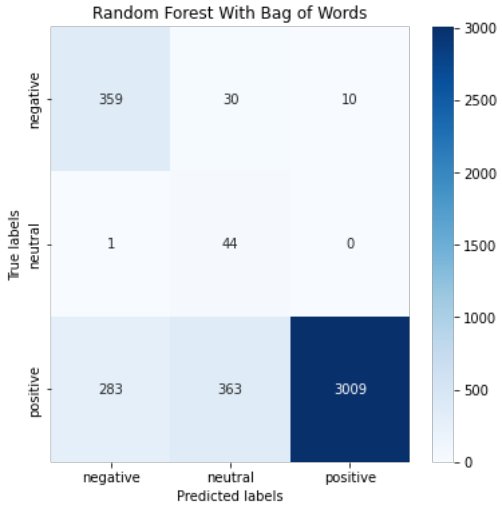


Figure 4.6: Confusion Matrix (Random Forest + Bag of Words)

4.2.2 Random Forest with TF-IDF

Figure 4.7 illustrates the confusion matrix for Random Forest using TF-IDF model for Datafiniti dataset, which indicates the relationship between the true labels and predicted labels. The total actual positive label for this model is 1458 from which 1414 are true positive labels and 44 are false negative labels. Total actual negative label is 542, from which 263 labels are predicted as true negative and 279 are counted as false positive.

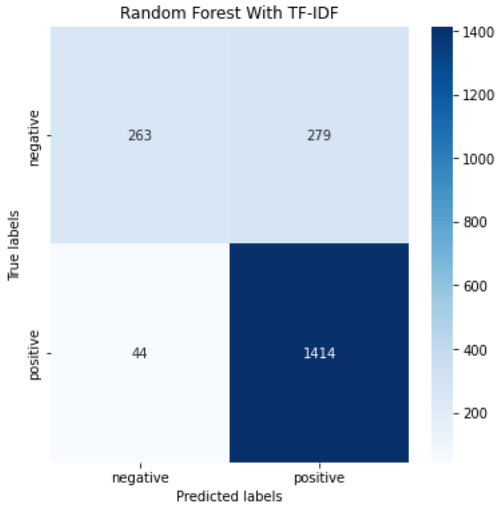


Figure 4.7: Confusion Matrix (Random Forest + TF-IDF)

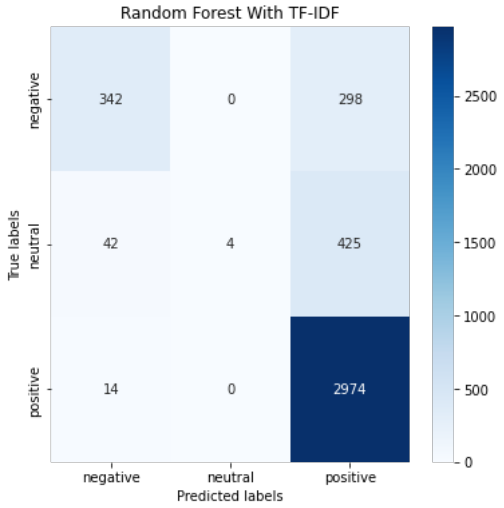


Figure 4.8: Confusion Matrix (Random Forest + TF-IDF)

Figure 4.8 represents the confusion matrix for Random Forest with TF-IDF on Tripadvisor dataset. The table shows the relation between true labels and predicted labels. Total number of actual positive label is 2988 from which 2974 are true positive (TP) labels, 0 label found as false neutral1 (FNt1) labels, and 14 are predicted false negative1 labels (FNg1). Total actual negative label is 640 where 342 labels are true negative (TNg), none of the labels is false neutral2 (FNt2) and 298 labels are counted as false positive1 (FP1). The total number of actual neutral is 471, among

that 42 labels are predicted as false negative2 (FNg2), 4 are true neutral (TNt) and 425 labels are found as false positive2 (FP2).

4.3 Support Vector Machine

4.3.1 SVM with Bag of Words

After testing the Datafiniti dataset in the Bag of Words model using the Support Vector Machine algorithm, the accuracy is 87.55 percent. Moreover, the scores of precision, recall, and F1-score are 87.79 percent, 87.55 percent 83.99 percent respectively. Figure 4.9 shows the confusion matrix for the Support Vector Machine

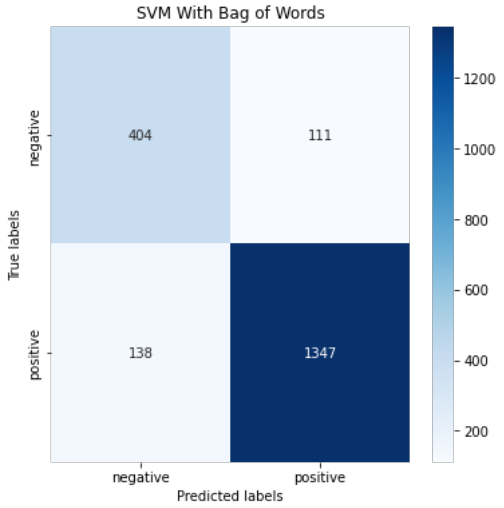


Figure 4.9: Confusion Matrix (SVM + Bag of Words)

with Bag of Words model for the Datafiniti dataset. It shows the relationship between true labels and predicted labels. The total actual positive is 1485 where 1347 are true positives and 138 are false negatives. Then, the total actual negative is 515. It represents the true negative and false positive scores. The score of the true negative is 404 and the false positive is 111. For the Tripadvisor dataset, SVM has the accuracy of 83.55 percent by using the Bag of Words model. Furthermore, the precision score is 85.23 percent and recall is 83.55 percent. Lastly, the F1-score is 68.47 percent.
Figure 4.10 shows the confusion matrix for the Support Vector Machine with Bag of Words model for Tripadvisor dataset. The figure also shows the relationship between true labels and predicted labels. The total actual negative is 331. True negative (TNg) is 310 and false positive1 score (FP1) is 3. But the false neutral2 score (FNt2) is 18. Furthermore, the total actual neutral score is 679. The false negative2 score (FNg2) is 255 and false positive2 score (FP2) is 147. True neutral (TNt) score is 277. Lastly, the total positive score is 3089. The true positive (TP) score is, 2838 and the false negative1 (FNg1) score is 75. Lastly, the false neutral1 (FNt1) score is 176.

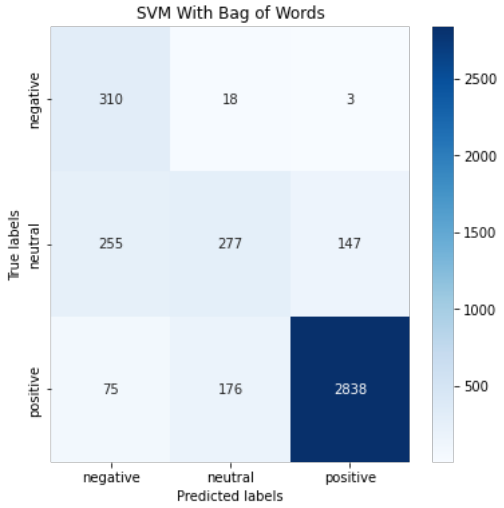


Figure 4.10: Confusion Matrix (SVM + Bag of Words)

4.3.2 SVM with TF-IDF

For Datafiniti dataset and TF-IDF model, Support Vector Machine algorithm has performed well. For this technique, the accuracy is 88.9 percent. Furthermore, SVM with TF-IDF model has a precision score of 89.75 percent and the recall score is 88.9 percent. Lastly, the F1-score is 85.27 percent.

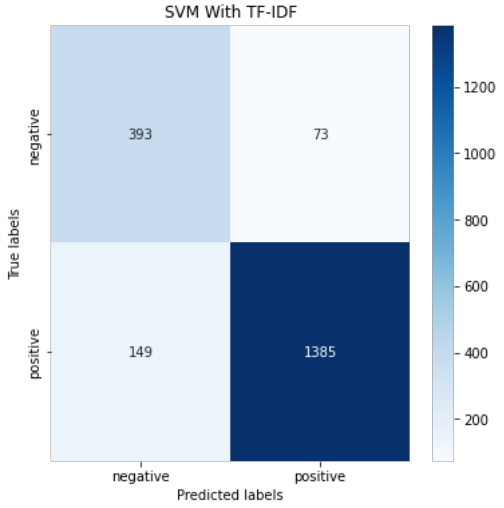


Figure 4.11: Confusion Matrix (SVM + TF-IDF)

Figure 4.11 demonstrates the confusion matrix for the Support Vector Machine with TF-IDF model for the Datafiniti dataset. The figure shows the relationship between true labels and predicted labels. The total actual positive is 1534. It is shown that, 1385 are true positives and 149 are false negatives. Moreover, the total actual negative is 466. True negative and false positive scores are represented by the total actual negative score. It shows that the score of the true negative is 393 and the false positive is 73.
By using the TF-IDF vectorization model for the Tripadvisor dataset, the accuracy is 89.24 percent. Precision score, recall score and F1- scores are 92.83 percent, 89.24 percent and 76.07 percent respectively.

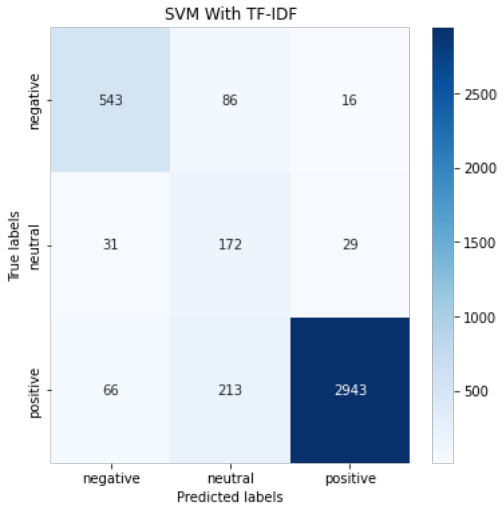


Figure 4.12: Confusion Matrix (SVM + TF-IDF)

Figure 4.12 shows the confusion matrix for the Support Vector Machine with TF-IDF model for the Tripadvisor dataset. It also shows the relationship between true labels and predicted labels. The total actual negative score is 645. True negative (TNg) score is 543 and false positive1 score (FP1) is 16. But the false neutral2 score (FNt2) is 86. Then, the total actual neutral score is 232. So we get the false negative2 score (FNg2) which is 31, total neutral (TNt) score 172, and false positive2 score (FP2) is 29. Lastly, total actual positive is 3222. Here, the true positive (TP) is 2943. False neutral1 (FNt1) is 213 and false negative1 (FNg1) is 66.

4.4 Logistic Regression

4.4.1 Logistic Regression with Bag of Words

After testing the Datafiniti dataset with Bag of Words model using the Logistic Regression algorithm, 89.15 percent accuracy and 85.65 percent F1-score was achieved.

Figure 4.13 shows the confusion matrix for the Logistic Regression with Bag of Words model for the Datafiniti dataset. The figure illustrates the true labels and predicted labels. The total actual positive is 1529. Between them, 1385 are true positives, and 144 are false negatives. As well as the total actual negative is 471. The score of the true negative is 398 and the false positive is 73.
While with Tripadvisor dataset with the Bag of Words model using the Logistic Regression algorithm, it provided 89.77 percent accuracy with 80.22 percent F1-score.

Figure 4.14 shows the confusion matrix for Logistic Regression with the Bag of Words model for the Tripadvisor dataset. The total actual negative score is 649.

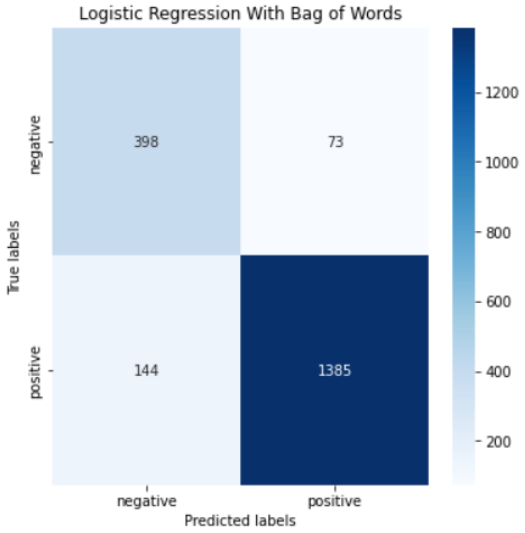


Figure 4.13: Confusion Matrix (Logistic Regression + Bag of Words)

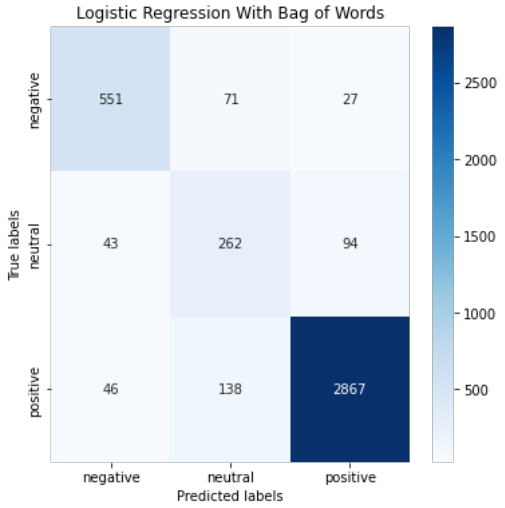


Figure 4.14: Confusion Matrix (Logistic Regression + Bag of Words)

The true negative (TNg) score is 551 and the false positive1 score (FP1) is 27. But the false neutral2 score (FNt2) is 71. Furthermore, the total actual neutral score is 399. So we get the false negative2 score (FNg2) which is 43, true neutral (TNt) score 262, and the false positive2 score (FP2) is 94. Lastly, total actual positive is 3051. Here, the true positive (TP) is 2867. False neutral1 (FNt1) is 138 and false negative1 (FNg1) is 46.

4.4.2 Logistic Regression with TF-IDF

After testing the Datafiniti dataset in the TF-IDF vectorization technique using a Logistic Regression algorithm, the model obtained 86.85 percent accuracy, and 81.78 percent F1-score.

Figure 4.15 shows the confusion matrix for the Logistic Regression algorithm with the TF-IDF vectorization technique for the Datafiniti dataset. The figure displays

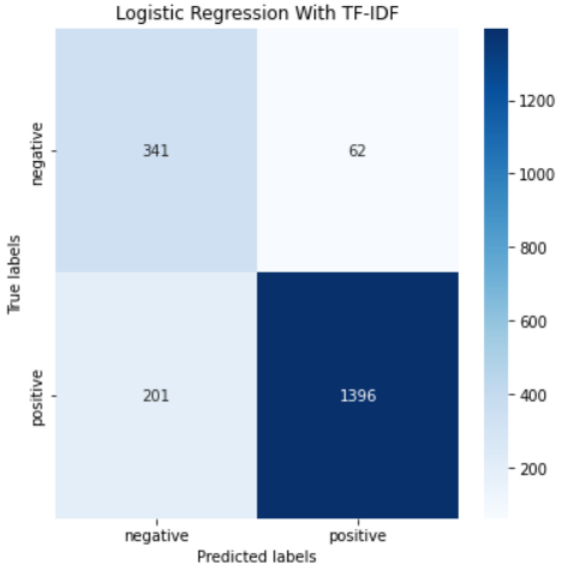


Figure 4.15: Confusion Matrix (Logistic Regression + TF-IDF)

the relationship between true labels as well as the estimated labels. The total actual positive is 1597. Between them, 1396 are true positives, and 201 are false negatives. In addition, the total actual negative is 403. It displays the true negative and false positive scores. The true negative is 341 and the false positive is 62.
As for the Tripadvisor dataset with TF-IDF vectorization technique using a Logistic Regression algorithm, the model obtained 87.75 percent accuracy and 72.28 percent F1-score.

Figure 4.16 displays the true labels as well as the predicted labels. The total

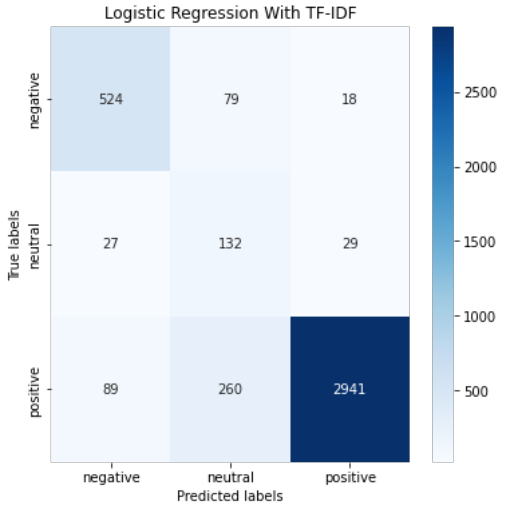


Figure 4.16: Confusion Matrix (Logistic Regression + TF-IDF)

actual negative score is 621. The true negative (TNg) score is 524 and the false positive score (FP1) is 18. But the false neutral2 score (FNt2) is 79. Then, the total actual neutral score is 188. So we get the false negative2 score (FNg2) which is 27 and the false positive2 score (FP2) is 29. The true neutral (TNt) score is 132.

Lastly, the total positive score is 3290. The true positive (TP) score is 2941 and the false negative1 (FNg1) score is 89. The false neutral1 (FNt1) score of 260.

4.5 Result Comparison

A research paper [11] shows their accuracy for four classifiers using TF-IDF and Word2Vec embedding process. For polarity-based approach, their best outcome is obtained from SVM model. This model generates 78% accuracy with TF-IDF and 81% accuracy for Word2Vec. Other classification models: Naïve Bayes, Logistic Regression, and Random Forest has a range of accuracy from 73% to 75% for TF-IDF and 74% to 80% for Word2Vec.
In this study, the accuracy range is 73.31% - 89.24% using TF-IDF and 73.23% - 89.77% for Bag of Words. For TF-IDF, SVM algorithm outperforms other models, and in case of BoW, Logistic Regression has the best outcome. Naïve Bayes has the lowest score for this dataset. The accuracy range is 76.7% - 88.9% (TF-IDF), and 77.9% - 89.15% (BoW) for the Datafiniti dataset. Same as the Tripadvisor dataset, SVM has the best accuracy in TF-IDF, Logistic Regression generated the highest accuracy in BoW. Naïve Bayes still underperforms other algorithms with the lowest accuracy.
Another study [21] is conducted using Multinomial Naïve Bayes, Random Forest, and SVM model with two parameters for each model. Here, Multinomial Naïve Bayes with alpha=0.009 has the highest accuracy (83.90%) with the best F1-score (67.16%). In our research, for both datasets, SVM has the highest accuracy with highest F1-score (Datafiniti: 85.27%, Tripadvisor: 76.07%) for TF-IDF. For BoW model, Logistic Regression has the best accuracy with F1-score (Datafiniti: 85.65%, Tripadvisor: 80.22%).

4.5.1 ROC Curve and AUC

Figure 4.17 shows the ROC Curve and AUC for two most effective models in this study. AUC for SVM with TF-IDF is 0.83, on the other hand, AUC for Logistic Regression with BoW is 0.79. Area under the curve for SVM is greater, which means it can predict positive and negative sentiments better than Logistic Regression.

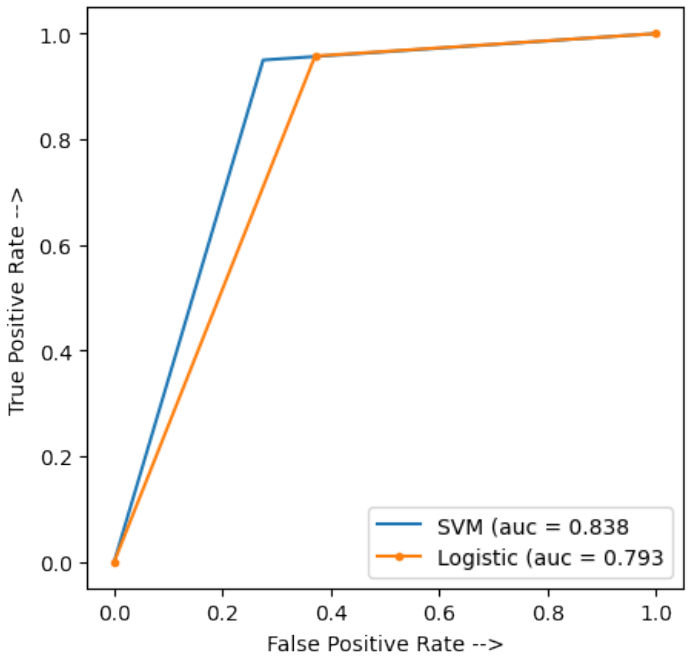


Figure 4.17: ROC Curve and AUC (SVM vs Logistic Regression for Datafiniti)

Chapter 5

Conclusion

5.1 Conclusion

Every opinion given online matters when it comes to the hotel business. Customer feedback has a significant impact on the hotel's reputation, which will help it grow its clientele in the future. In this paper, a system to analyze the efficiency using four different classification algorithms has been proposed. Two datasets from Datafiniti and Tripadvisor have been used. For these two different datasets, different performances for each algorithm have been found. It was possible to get almost 73-90% accuracy using the four classifiers for the datasets. For both datasets, SVM and Logistic Regression outperforms the other two algorithms. In the proposed thesis, a generalized framework for customers and hotels to make decisions based on sentiments has been built.

5.2 Future Work

This research work is expected to help pave the path to building a recommender system for the customers in order to facilitate the hotels to monitor and upgrade their services accordingly. It will help them to make their decision more effectively and accurately. Usage of additional classifiers with balanced datasets can be implemented to get a much more efficient model in the future for better improvement of the cause.

Bibliography

[1] E. Loper and S. Bird, "Nltk: The natural language toolkit," *CoRR*, vol. cs.CL/0205028, 2002. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr0205.html#cs-CL-0205028.

[2] P. Stalidis, "Use of machine learning algorithms for sentiment analysis in online hotel reviews," Ph.D. dissertation, Jun. 2015.

[3] T. Pranckevicius and V. Marcinkevičius, "Application of logistic regression with part-of-the-speech tagging for multi-class text classification," Nov. 2016, pp. 1–5. DOI: 10.1109/AIEEE.2016.7821805.

[4] S. Rezaeinia, A. Ghodsi, and R. Rahmani, "Improving the accuracy of pre-trained word embeddings for sentiment analysis," Nov. 2017.

[5] Shahnawaz and P. Astya, "Sentiment analysis: Approaches and open issues," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, pp. 154–158. DOI: 10.1109/CCAA.2017.8229791.

[6] B. Bansal, *Tripadvisor hotel review dataset*, Zenodo, Apr. 2018. DOI: 10.5281/zenodo.1219899. [Online]. Available: https://doi.org/10.5281/zenodo.1219899.

[7] G. Khanvilkar and D. Vora, *Sentiment analysis for product recommendation using random forest*, Jun. 2018. DOI: 10.14419/ijet.v7i3.3.14492.

[8] K. Zvarevashe and O. O. Olugbara, "A framework for sentiment analysis with opinion mining of hotel reviews," in *2018 Conference on Information Communications Technology and Society (ICTAS)*, 2018, pp. 1–4. DOI: 10.1109/ICTAS.2018.8368746.

[9] B. Bahrawi, "Sentiment analysis using random forest algorithm online social media based," vol. 2, h.29–33, Dec. 2019.

[10] Datafiniti, *Hotel reviews*, Jun. 2019. [Online]. Available: https://www.kaggle.com/datasets/datafiniti/hotel-reviews.

[11] S. Abro, S. Shaikh, R. A. Abro, S. F. Soomro, and H. M. Malik, "Aspect based sentimental analysis of hotel reviews: A comparative study," *Sukkur IBA Journal of Computing and Mathematical Sciences*, vol. 4, no. 1, pp. 11–20, 2020.

[12] M. Beny, A. Barakbah, and T. Muliawati, "Data analytics for hotel reviews in multi-language based on factor aggregation of sentiment polarization," Sep. 2020, pp. 324–331. DOI: 10.1109/IES50839.2020.9231625.

[13] V. Chang, L. Liu, Q. Xu, T. Li, and C.-H. Hsu, "An improved model for sentiment analysis on luxury hotel review," *Expert Systems*, e12580, 2020.

[14] M. IŞIK and H. Dağ, "The impact of text preprocessing on the prediction of review ratings," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 3, pp. 1405–1421, 2020.

[15] N. Kasture, *Why it is important to handle missing data and 10 methods to do it.* Jul. 2020. [Online]. Available: https://medium.com/analytics-vidhya/why-it-is-important-to-handle-missing-data-and-10-methods-to-do-it-29d32ec4e6a.

[16] S. Khomsah, "Naive bayes classifier optimization on sentiment analysis of hotel reviews," *Jurnal Penelitian Pos dan Informatika*, vol. 10, p. 157, Dec. 2020. DOI: 10.17933/jppi.2020.100206.

[17] K. Lal and N. Mishra, "Feature based opinion mining on hotel reviews using deep learning," in Jan. 2020, pp. 616–625, ISBN: 978-3-030-38039-7. DOI: 10.1007/978-3-030-38040-3_70.

[18] X. Li and C. Liu, "Comparison of machine learning models for sentimental analysis of hotel reviews," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 806, 2020, p. 012 029.

[19] M. McGregor, *Svm machine learning tutorial – what is the support vector machine algorithm, explained with code examples*, Jul. 2020. [Online]. Available: https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/.

[20] S. Zahid, *Sentiment analysis of hotel reviews - performance evaluation of machine learning algorithms*, Jan. 2020. DOI: 10.13140/RG.2.2.21026.96965.

[21] S. Zahid-samza595, "Sentiment analysis of hotel reviews-performance evaluation of machine learning algorithms," 2020.

[22] S. Jayalal, "Analysis of helpfulness of online hotel reviews: Classifica- tion based approach," Dec. 2021.

[23] C. Khanna, *Text pre-processing: Stop words removal using different libraries*, Feb. 2021. [Online]. Available: https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a.

[24] S. Kumar, *7 ways to handle missing values in machine learning*, Sep. 2021. [Online]. Available: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e.

[25] A. Saini, *Naive bayes algorithm: A complete guide for data science enthusiasts*, Sep. 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/09/naive-bayes-algorithm-a-complete-guide-for-data-science-enthusiasts/.

[26] R. Team, *Sentiment analysis challenges: Everything you need to know*, Jun. 2021. [Online]. Available: https://www.repustate.com/blog/sentiment-analysis-challenges-with-solutions/.

[27] G. Lawton, E. Burns, and L. Rosencrance, *What is logistic regression? - definition from searchbusinessanalytics*, Jan. 2022. [Online]. Available: https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression.

[28] Saumyab271, *Stemming vs lemmatization in nlp: Must-know differences*, Jul. 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/#:~:text=Stemming%20is%20a%20process%20that,'%20would%20return%20'Car'.

[29] J. Selig, *What is machine learning? a definition.* Jul. 2022. [Online]. Available: https://www.expert.ai/blog/machine-learning-definition/.

[30] [Online]. Available: https://www.prnewswire.com/news-releases/online-reviews-remain-a-trusted-source-of-information-when-booking-trips-reveals-new-research-300885097.html?fbclid=IwAR06SiClzFn3XjBCecxvOvX-QejLAJlot66Iy6RkIc_oyfAeN5iaLEutxJs.

[31] *Naive bayes classifier in machine learning - javatpoint.* [Online]. Available: https://www.javatpoint.com/machine-learning-naive-bayes-classifier.

[32] *Support vector machine (svm) algorithm - javatpoint.* [Online]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[33] *Text classification: What it is how to get started.* [Online]. Available: https://levity.ai/blog/text-classification.

[34] *Unsupervised machine learning - javatpoint.* [Online]. Available: https://www.javatpoint.com/unsupervised-machine-learning.