

# Aspect Based Opinion Mining on Restaurant Reviews

by

Rafi Ahmed

16101065

Sazid Hasan Tonmoy

16301003

Mehejabin Binta Bashar

18101568

Madhurjya Sarkar

18101574

Faiyaj Bin Ahmed

23141084

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

Department of Computer Science and Engineering  
School of Data and Sciences  
Brac University  
May 2023

© 2023. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

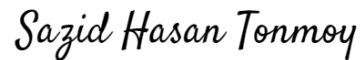
1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material 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:**



---

Rafi Ahmed  
16101065



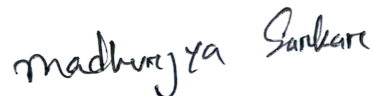
---

Sazid Hasan Tonmoy  
16301003



---

Mehejabin Binta Bashar  
18101568



---

Madhurjya Sarkar  
18101574



---

Faiyaj Bin Ahmed  
23141084

# Approval

The thesis titled “Aspect Based Opinion Mining on Restaurant Reviews” submitted by

1. Rafi Ahmed(16101065)
2. Sazid Hasan Tonmoy(16301003)
3. Mehejabin Binta Bashar (18101568)
4. Madhurjya Sarkar(18101574)
5. Faiyaj Bin Ahmed(23141084)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 17, 2023.

## Examining Committee:

Supervisor:  
(Member)



---

Farig Yousuf Sadeque  
Associate Professor  
Department of Computer Science and Engineering  
Brac University

Co-Supervisor:  
(Member)



---

Moin Mostakim  
Senior Lecturer  
Department of Computer Science and Engineering  
Brac University

Program Coordinator:  
(Member)

---

name here  
Professor  
Department Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)

---

Dr. Sadia Hamid Kazi  
Chairperson and Associate Professor  
Department Computer Science and Engineering  
Brac University

## **Ethics Statement (Optional)**

This is optional, if you don't have an ethics statement then omit this page

# Abstract

The way businesses are operating have changed due to the explosion of the internet. Social media has an increasing number of reviews as people are keen to express their opinions based on their experiences. Online reviews have become a precious asset in various disciplines such as intelligent marketing and decision-making. The number of reviews for a well-liked product might reach thousands. This makes it challenging for a prospective buyer to go through them and make up their minds. In order to overcome this challenge, a machine-learning system is needed. Aspect based Opinion mining can be used to extract the aspects from the reviews, then we can analyze the nature of the reviews and recommend them to all the customers. We plan to classify reviews about a target entity as positive, negative and neutral so that readers of the reviews do not have to go through all the reviews but instead can focus on functional items and applicable suggestions. This thesis is specifically focused on reviews in the domain of restaurants. This study extends our knowledge of online reviews by taking into account users' wants and anticipating their future behavior. Several distinct evaluative linguistic nuances shed light on internet reviews. Using an assortment of models on generated benchmark datasets, we will also empirically show the efficacy of our strategy and show that the new techniques (or modified versions) are superior to, or at least on par with, state-of-the-art methods.

**Keywords:** Extract aspects; Online reviews; Automated methodology; Customers; Analyze nature

## Dedication (Optional)

A dedication is the expression of friendly connection or thanks by the author towards another person. It can occupy one or multiple lines depending on its importance. You can remove this page if you want.

# Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor and co-supervisor for their kind support and advice in our work. They helped us whenever we needed help.



# Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	v
Dedication	vi
Acknowledgment	vii
Table of Contents	viii
List of Figures	x
List of Tables	1
<b>1 Introduction</b>	<b>2</b>
1.1 Research Problem . . . . .	2
1.2 Research Objectives . . . . .	4
<b>2 Literature Review</b>	<b>5</b>
2.1 Related Works . . . . .	6
<b>3 Methodology</b>	<b>8</b>
3.1 Data Collection . . . . .	10
3.2 Data Pre-Processing . . . . .	10
3.2.1 Tokenization . . . . .	10
3.2.2 Stop Words Removal . . . . .	11
3.2.3 Removal of rare words . . . . .	11
3.2.4 Stemming . . . . .	18
3.2.5 Lemmatization . . . . .	18
3.2.6 Conversion of Emoji to Words . . . . .	19
3.2.7 Removal of emojis . . . . .	19
3.2.8 Removal of URLs . . . . .	20
3.2.9 Part of Speech Tagging . . . . .	20
3.3 Topic Modelling . . . . .	20
3.3.1 Latent Dirichlet Allocation . . . . .	21

3.4	Classifications . . . . .	23
3.4.1	Multinomial Naive Bayes . . . . .	23
3.4.2	Random Forest . . . . .	23
3.4.3	Support Vector Machine . . . . .	24
<b>4</b>	<b>Result &amp; Analysis</b>	<b>26</b>
4.0.1	Confusion Matrix for Naive Bayes . . . . .	26
4.0.2	Confusion Matrix for SVM . . . . .	27
4.0.3	Confusion Matrix for Random Forest . . . . .	27
<b>5</b>	<b>Conclusion</b>	<b>29</b>
	<b>Bibliography</b>	<b>31</b>

# List of Figures

3.1	Flow chart of the aspect extraction model. . . . .	9
3.2	Dataset before pre-processing. . . . .	10
3.3	Dataset after pre-processing. . . . .	10
3.4	rating_review, review_len, word_count, polarity. . . . .	11
3.5	Visualization of text data using wordcloud. . . . .	12
3.6	Histogram of review_len, word_count and polarity. . . . .	12
3.7	Number of data each rating. . . . .	13
3.8	Box plot of review rating of vs. polarity. . . . .	13
3.9	Point plot of product rating vs review length. . . . .	14
3.10	Bar chart of top word frequency. . . . .	14
3.11	Bar chart of bigram frequency. . . . .	15
3.12	Stop words removal process. . . . .	15
3.13	Top 20 unigrams before removing stopwords. . . . .	16
3.14	Top 20 unigrams after removing stopwords. . . . .	16
3.15	Top 20 bigrams before removing stopwords. . . . .	17
3.16	Top 20 bigrams after removing stopwords. . . . .	17
3.17	Process of Stemming. . . . .	18
3.18	Process of Lemmatization. . . . .	19
3.19	Part of speech tagging of review text. . . . .	20
3.20	Keyword extracted using LDA (part1). . . . .	22
3.21	Keyword extracted using LDA (part2). . . . .	22
3.22	Random forest algorithm. . . . .	24
4.1	Confusion Matrix. . . . .	26
4.2	Confusion Matrix for Naive Bayes . . . . .	27
4.3	Confusion Matrix for SVM . . . . .	27
4.4	Confusion Matrix for Random Forest . . . . .	28

# List of Tables

4.1	Result analysis . . . . .	26
-----	---------------------------	----

# Chapter 1

## Introduction

The online review has become very essential as it became a predestined part of consumers' decision-making process in going to a restaurant. A huge amount of people admitted that they "consistently or occasionally" look up the restaurant review in advance of going to a restaurant. Although the customer review is increasing in a large quantity it varies largely in quality. In the present time, a restaurant can get a hundred to thousands of reviews in a short span of time but within this amount of reviews, the high-quality reviews get mixed up with the vague or useless ones. So for the people who are interested in coming to a restaurant, it would be really hard and tiresome to go over these reviews to make a decision. To solve this problem, Opinion mining is used. It is a field that scrutinizes people's thoughts, emotions, and expressions toward a restaurant based on its food quality, hygiene and staff behavior, cost, etc. It basically talks about the positive and negative sentiments or the magnitude of the criticism or the praise regarding the restaurant or entities of the restaurant. The sentiments demonstrated in the review can be researched at different levels of granularity. Like Document-level opinion looks upon allocating total sentiment over the whole review. Moreover, Sentence-level opinion searches the sentiments in each of the sentences in the review. Furthermore, there is an opinion level where the analysis is done on every word in the reviews as the polarity can be different in different situations. Aspects point to unique dimensions of an entity and also it can be a semantic function or a specific feature. For example "The size of the restaurant is big and the interior design is very nice. And I recommend it though the price of the food is quite extravagant." In this review, we can see the overall review seems to be good but there is conflicting opinion between different aspects like the size and the design as constructive opinions while the polarity of the "price" was negative which can be seen to be indicated by the word "extravagant".

### 1.1 Research Problem

With the increase of restaurants all over the world, it is very hard to maintain communication and relationships with customers. The most complicated challenge with the growing number of customers is the large amount of data that is generated in the form of natural language. Extracting all the necessary information from this data is very difficult as there exist different meanings to a sentence. To make this

data useful there are major challenges to solve in aspect-based opinion mining.

Aspect extraction comes first. This task entails identifying the traits and characteristics of a target item from the review. The key objective of this work is to ascertain which particular features of a reviewed item (product) are considered important in customer evaluations. Whenever extracting an aspect it is very important to know which entity this belongs to. Aspect can be said to be divided into two parts. The first one is the Explicit aspect. When the aspect is in the sentence for instance, when we say “The Beef Burger tastes awesome” here “taste” is the aspect which is used in the sentence explicitly. The second one is the Implicit aspect. In the sentence “The restaurant is small” here talking about the size of the restaurant but it is not mentioned directly. In the aspect extraction, there are some things to work with such as searching the type of the aspect for opinion mining and selecting good aspects for classification.

The Aspect Opinion Classification is the second. It is integral to categorize the opinion words that were extracted as well as those associated with the multiple traits into one of the three polarity scales (positive, negative, or neutral). It strives to ascertain the specific numerical opinion evaluations on the aspects or determine if interpretations and feelings toward the identified aspects are positive, negative, or neutral. This Opinion classification can be performed in the form of stars, thumbs-up, or thumbs-down. For example, “iPhone camera quality is amazing”, the aspect opinion classification task will find the opinion orientation of the word “amazing” on the “camera quality” aspect as positive. Different algorithms like n-gram, naive Bayes, vector semantics, and logistic regression are used to classify the extracted aspects.

Thirdly, Finding implicit aspects is a very challenging task as people express their opinion differently and the habits of language differ from person to person. Many reviews refer to multiple aspects of different products. For example, “The restaurant is very much smaller compared to another one”, here ‘The restaurant is very much smaller’ tells us about the restaurant’s ‘size’ aspect. When a single implicit feature is found then we can quite easily detect the sentiment of that feature but when there is more than one implicit feature present then detecting the sentiment of the feature becomes more challenging. In case of multi implicit features, the features may have a different level of polarities than the complete polarity of the sentence. For example, in the statement “Pictures taken can get blurred because of lack of image stabilizer but overall a great option for a given budget”, two different aspects camera quality and price are mentioned implicitly. Consider this example: “Food quality of the restaurant is very good but it is not that tidy”. In this review, the restaurant’s food quality and hygiene are brought up. Since people have positive and negative thoughts about many things, it is wise to resolve the numerous sorts of elements and take cognizance of their polarity autonomously. Finding several facets of a feature is thus a challenging endeavor.

The fourth is the Cross-Domain adaptation. The majority of opinion mining systems heavily rely on certain domains. The same perspective phrase might convey various polarities in other contexts. Prior to mining opinions from the reviews, subject understanding is imperative. It may prove daunting for academics to generate domain-independent methodologies and algorithms

Next is modeling the customer's review with a view of finding semantics aspects and opinions and also projecting the overall rating review. Usually, the customer gives a review with an overall rating in the form of stars. So by utilizing the overall rating it is possible to lead the process of the sentiment aspect of the review.

## **1.2 Research Objectives**

The objective of the research is to improve the effectiveness of aspect extraction. Different types of techniques can be used to search, extract and look up relevant information. The main objectives of our research are given below :

1. Form an effective model to search and extract all the aspects using NLP techniques.
2. Perform sentiment analysis on all the aspects extracted from the reviews.
3. Identify a mapping between the opinions and the extracted aspects.

# Chapter 2

## Literature Review

For decision making opinions can play a major role to choose from multiple choices involving valuable resources. Until recently, friends and specialized magazines or websites were the main sources of information. Our ability to easily produce and exchange ideas with everyone linked through discussion boards, blogs, social media platforms, and content-sharing services has evolved thanks to the internet, which has given us an abundance of possibilities and new tools. According to Statista [10], in 2020, social media users reached over 3.6 billion people worldwide, a number projected to increase to almost 4.41 billion in 2025. However, a new study by Kepios claims that by April 2022, there will be approximately 4.65 billion online social networking users globally, which is 58.7% of the whole of humanity's populace. . Due to the epidemic, social media engagement rates have surged over the previous 12 months as well, with 326 million new individuals joining during this period last year. This translates to a yearly increase of 7.5 percent, or in excess of 10 new users per second on average. [15]. As a result the number of shared opinions on various topics over the internet is exponentially increasing. Due to the unstructured nature of this information it's not machine processable. The scientific community is becoming increasingly proactive in gathering public opinion on many issues because of the possible difficulties that may arise from exploring uncharted areas. The mining of opinions and sentiment is one of the new areas that have emerged as a consequence of this. Aspect Based Opinion Mining(ABOM) aims to extract aspects of products and classify the corresponding polarities of the user in the review. Previously, several approaches have been used to study ABOM based on text reviews. Vector extraction is now utilized in opinion extraction and sentiment evaluation to acquire the most prominent and significant linguistic properties. Frequency and presence are the two most common features in Vector classification [9]. Additionally, n-grams, which are often bigrams and trigrams, are seen as having useful qualities. While linguistic evaluation employs part of speech (also known as POS) information, such as adjectives, nouns, adverbs, and verbs, as a fundamental kind of Word-Sense Disambiguation (WSD), other approaches additionally rely on the disparity within words. The aforementioned techniques are firmly constrained by topic and domain.[1].



## 2.1 Related Works

In the framework of aspects-based sentiment analysis, this section tries to critically analyze prior pertinent research in the field of opinion mining [15]. We looked at the many methods and strategies applied to get the desired result. Opinion mining encompasses data mining, computational linguistics, and natural language processing (NLP). Unlike its predecessor, normal syntactical NLP, opinion mining doesn't require a thorough grasp of the text. While Syntactical NLP emphasizes on summarization and auto-classification categorizing, it primarily focuses on semantic inference and emotional information related to natural language [15]. Opinion mining may be used at several levels, including the document, phrase, entity, and aspect levels. Whether a whole opinion paper reflects a favorable or negative mood is dependent on the document level. While sentence-level analyzes each sentence to determine if it is neutral, unfavorable, or expresses an opinion. Since entity and aspect level analysis looks at the viewpoint itself rather than linguistic structures (documents, paragraphs, sentences, clauses, or phrases), it is more precise [10]. It is based on the idea that an opinion is made up of a subject, a feeling (whether positive or negative), and both.. Aspect identification, ABM word identification, and ABM word orientation detection are the three main tasks in ABM [3]. Think of a restaurant review that says, "Although the food is good, the service is not that great." The terms "not great" and "good" are used to express opinions about the features of service and cuisine in this context.

Using movie reviews as their data, the researchers in this work [10] categorised papers by overall sentiment rather than themes and discovered that typical machine learning algorithms surpass baselines created by humans. They used three machine learning methods: Naive Bayes, maximum entropy classification, and support vector machines. These methods are excellent at topic-based categorization but poor at classifying sentiment. Aspect word extraction, aspect category identification, and aspect sentiment prediction are three of the subtasks., Alghunaim et al. [10] have examined the efficacy of vector representations over various text data. When compared to the baselines, the vector space method utilized in this work scored well, with F1 scores of 79.91. By using a hierarchical bidirectional LSTM to describe the relationships between phrases in a review, the authors of [2] claimed to have proven the job of aspect-based sentiment analysis. The suggested hierarchical model beats two nonhierarchical baselines, achieving results on par with state-of-the-art datasets and even outperforming them on a few multilingual, multi-domain datasets without the need of hand-engineered features or outside resources. A technique called Sentiment Utility Logistic Model (SULM) was suggested in this research [6] for determining the most useful elements of upcoming user experiences. When tested on genuine evaluations from three real-world apps, our technique functioned admirably. Additionally, equivalent to the most advanced HFT model, It was able to predict the 6-level unknown ratings of the reviews. In addition, it predicted the group of components that the user would mention in a prospective future assessment of an item at the level of the most sophisticated LRPPM. Service providers can gain from the beneficial user experience characteristics offered by SULM to assure better services for users. The research study [7] offered the initial deep learning approach to opinion mining aspect extraction, which involves finding opinion targets in the text that has been expressed with strong opinions. The proposed deep CNN archi-

texture consists of seven layers: Two convolution layers, followed by a max-pooling layer following each, a fully connected layer, and then a layer of output with one neuron for each word are all integrated to create the final outcome. Each word in the sentence has word embedding features in the input layers. The neural network classifier were paired with a set of heuristic patterns of language that had already been produced. Performance with this model was notably superior than that with cutting-edge techniques.

Despite being effective for Aspect Based Sentiment Analysis, supervised learning methods fall behind due to lack of fine-grained labeled data. To address this issue, a new domain adaptation paradigm was proposed in this paper called Cross-domain review generation (CDRG). Using the source-domain labeled review, CDRG can generate fine-grained annotation based target-domain reviews. Through rigorous experiments CDRG has proven to be more effective than state-of-the-art domain adaptation methods [4].

The researcher Sazzed [5] emphasized the importance of working with annotated datasets that have varied demography as sociocultural factors and demography plays a vital role in a user's attitude and preference. This paper first constructs a dataset that focuses on local demography. Without relying on any labeled data, they proposed hybrid methodology and managed to outperform the best lexicon-based and Machine Learning (ML) based classifier. By using two contrasting local and a global datasets, further investigation was carried out to find out the effect of demography over linguistic characteristics. This conducted experiment showed that user demography plays a crucial role in the linguistic aspects of reviews.

This paper [12] introduces DomBERT, an extension of general-purpose language model, BERT that will assimilate both relevant domain corpora and in-domain corpora. The purpose was to combine domain oriented language model with general purpose language model to reduce the resource requirement for traditional domain language model. Conducted experiment showed assuring results which further establishes DomBERT's usefulness in ABSA.

Scarcity of fine-grained labeled data has always been one of the biggest challenges for the ABSA tasks specially for supervised models. To extenuate this limitation and dependency on labeled data, researchers primarily used feature-based domain adaptation or instance-based domain adaptation. But both of these methods have their pros and cons. To unravel this limitation, researchers proposed a new method, an end-to-end framework called Unified Domain Adaptation (UDA) that will merge feature-based adaptation with instance-based adaptation for both tasks of cross-domain aspect extraction and cross-domain End2End ABSA. Conducted experiment achieved remarkable results on four benchmarks and it has proven to be a significant improvement over the existing state-of-the-art methods for both tasks [11].

# Chapter 3

## Methodology

Our main scheme is to determine the different evaluation categories such as POS OPINION, NEG OPINION, MIX OPINION, SUGGESTION, INTENTION, DESCRIPTION by applying machine learning and deep learning approaches. We have used a dataset from Kaggle where there are reviews from customers who have reviewed specific restaurants. In our dataset, there are around 2.7 million reviews from different countries. We annotated each sentence into various categories such as positive, negative, neutral, intention, description etc. A sentence is annotated as the minority category when it consists of multiple categories. Online reviews are not machine processable information and may contain various types of noise. Therefore, Cleaning and normalization is essential for the analysis. In the cleaning step all the punctuation marks are removed and in the Case folding step all uppercase letters were converted to lowercase letters. After the preprocessing stage, for aspect extraction we first found the POS tag of words in each sentence to identify nouns and noun phrases. After that aspects are chosen based on the frequency. For the Subjectivity and objectivity classification stage, we used various conventional methods for text classification such as Linear SVM, CNN or LSTM. We used these methods to identify the subjective sentences and discarded the objective sentences based on aspects and opinion words. In Aspect related Opinion words Identification phase, aspects related to opinion words are identified. For understanding the effectiveness of our proposed model, we evaluated its performance by measuring accuracy, precision, recall and error rate. This will show us how our proposed system performs compared to the existing state-of-the-art methods.

In the data preprocessing, there are a number of cleanings we have done. The Processes labels are 1. dropping the null values, 2. dropping the duplicate values, 3. Omitting not useful features, 4. using Label Encoder to convert the string values to an int value, 5. converting the Date Time format, 6. Making the review text lowercase, 7. Expanding Contractions, 8. Removing digits and words containing digits, 9. removing punctuations and extra spaces, 10. removing stopwords, 11. removing frequent words. There exist several steps followed by this research method. Apart from that, we are also going to talk about the algorithms that we have used in our model. As we know there are various types of algorithms are there such as Naïve Bayes, Decision Tree, Random Forests, Support Vector Machines (SVM), Long Short-Term Memory network (LSTM), Convolutional Neural Network (CNN)

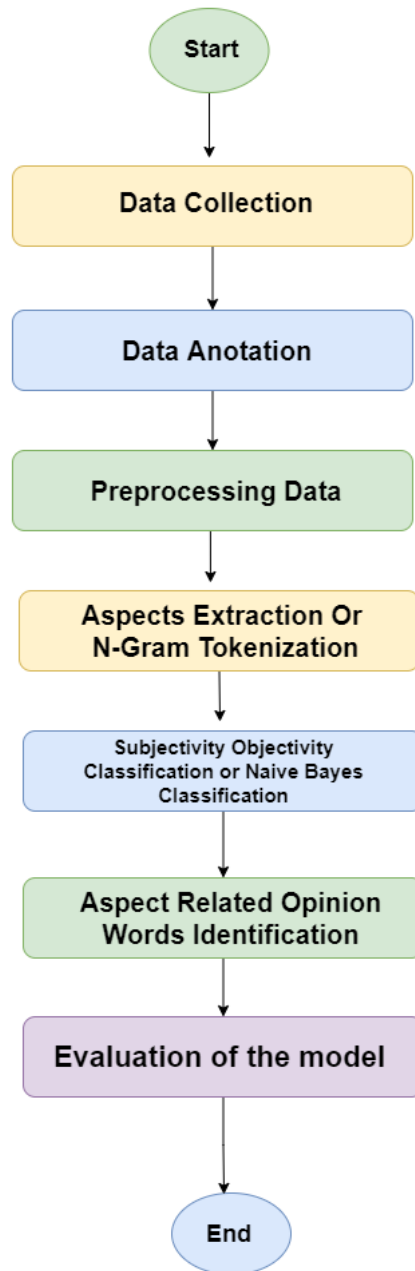


Figure 3.1: Flow chart of the aspect extraction model.

Unnamed: 0	parse_count	restaurant_name	rating_review	sample	review_id	title_review	review_preview	review_full	date	city	url_restaurant	author_id
612629	196818	The Queens Head	4.0	Positive	review_700475083	Authentic British pub in Piccadilly	This pub has two floors, ground floor is a pub...	This pub has two floors, ground floor is a pub...	August 18, 2019	London, England	https://www.tripadvisor.com/Restaurant_Review...	UID_145480
964835	550231	ROSSODISERA il marchigiano	5.0	Positive	review_45078306	A little piece of Italy in the heart of Covent...	We stumbled across this gem of a restaurant wh...	We stumbled across this gem of a restaurant wh...	January 9, 2017	London, England	https://www.tripadvisor.com/Restaurant_Review...	UID_327828
2207341	441815	Red Lobster	3	Negative	review_355023130	Good, but not great	While looking for a place to eat during out NY...	While looking for a place to eat during out NY...	March 15, 2016	New York City, New York	https://www.tripadvisor.com/Restaurant_Review...	UID_8898
238532	239161	El pebrot i el pellt_cargol	5	Positive	review_519641933	Awsome food and service at affordable prices!	We tried the rabbit stew with snails. It was a...	We tried the rabbit stew with snails. It was a...	August 30, 2017	Barcelona, Catalonia	https://www.tripadvisor.com/Restaurant_Review...	UID_156430
223636	224180	Art i Tapes	5	Positive	review_536122152	We loved it	Very nice food and staff. The Sangria was real...	Very nice food and staff. The Sangria was real...	October 26, 2017	Barcelona, Catalonia	https://www.tripadvisor.com/Restaurant_Review...	UID_39913
3507157	230211	Domaine de Lardin	3	Negative	review_17315415	Good but could have been better!	Our daughter had tried this restaurant before...	Our daughter had tried this restaurant before...	September 25, 2013	Paris, Ile de France	https://www.tripadvisor.com/Restaurant_Review...	UID_32653
68066	275072	108_Brasserie	4.0	Positive	review_211930389	Really Good	An all round good experience for business lanc...	An all round good experience for business lanc...	June 19, 2014	London, England	https://www.tripadvisor.com/Restaurant_Review...	UID_3603
2049194	281376	Arty_Ruth's Home_Style_Southern_Cuisine	5	Positive	review_42293844	Amazing Soul Food	This is Soul Food like 'Mama' makes. The food is...	This is Soul Food like 'Mama' makes. The food is...	September 27, 2016	New York City, New York	https://www.tripadvisor.com/Restaurant_Review...	UID_164112
2426410	144667	L'Appelo	5	Positive	review_7538881	Triumph of vegetables	Aline Pascard is a king of vegetables. Perfect...	Aline Pascard is a king of vegetables. Perfect...	August 15, 2010	Paris, Ile de France	https://www.tripadvisor.com/Restaurant_Review...	UID_101802
31536	317146	Los Canacoles	4	Positive	review_230172295	Anniversary meal	We were recommended this hotel so looked for a...	We were recommended this hotel so looked for a...	September 21, 2014	Barcelona, Catalonia	https://www.tripadvisor.com/Restaurant_Review...	UID_190361

Figure 3.2: Dataset before pre-processing.

parse_count	restaurant_name	rating_review	review_id	title_review	review_full	date	review_full_w_freqwords	review_full_w_removewords	review_full_lemmatized	review_len	word_count	polarity
612629	196819	The Queens Head	4.0	review_700475093	Authentic British pub in Piccadilly	August 18, 2019	pub two floors ground floor pub small restaura...	pub two floors ground floor pub small restaura...	pub two floor ground floor pub small restaura...	306	47	-0.167308
964835	550231	ROSSODISERA il marchigiano	5.0	review_45078306	A little piece of Italy in the heart of Covent...	January 9, 2017	stumbled across gem restaurant whilst weekend...	stumbled across gem restaurant whilst weekend...	stumbled across gem restaurant whilst weekend...	434	65	0.273333
2207341	441816	Red Lobster	3	review_355023130	Good, but not great	March 15, 2016	looking place eat my tip found outsdelf red l...	looking place eat my tip found outsdelf red l...	looking place eat my tip found outsdelf red l...	405	64	0.097222
238532	239161	El pebrot i el pellt_cargol	5	review_519641933	Awsome food and service at affordable prices!	August 30, 2017	tried rabbit stew snails absolutely delicious...	tried rabbit stew snails absolutely delicious...	tried rabbit stew snail absolutely delicious €...	122	17	0.712000
223636	224180	Art i Tapes	5	review_536122152	We loved it	October 26, 2017	nice food staff sangria really impressive enjo...	nice food staff sangria really impressive enjo...	nice food staff sangria really impressive enjo...	113	14	0.700000

Figure 3.3: Dataset after pre-processing.

etc. The following paper is showing all the step of this research paper.

## 3.1 Data Collection

Finding the labeled data for the text is very hard. There are multiple ways of extracting the data and one of the ways is by web scraping. But we have used a dataset from Kaggle where there are reviews from customers who have reviewed specific restaurants [3]. In our dataset, there are around 2.7 million reviews from different countries. Finding fine-grained labeled dataset for large text corpus is extremely hard. The dataset we used was taken from Kaggle under the name of “Six TripAdvisor Datasets for NLP Tasks” related to the paper “Explain and Conquer: Personalised Text-based Reviews to Achieve Transparency”[13]. It has 2.7M rows with 13 columns. All the columns are object types. This dataset contains restaurant reviews from six cities across the world, Barcelona, London, Madrid, New Delhi, New York and Paris. An excerpt from the dataset before cleaning can be seen here in Fig-

## 3.2 Data Pre-Processing

### 3.2.1 Tokenization

Tokenization is the focus of the conversion of sentences into coherent bits of data so that a program can work with. Using tokenization we broke the raw review text into words and sentences. These small chunks are known as tokens which aid in context comprehension or model development for Natural language processing. By examining the order of words in the text, tokenization helps in comprehending the text’s meaning. For example, ‘Love this place’ can be converted into ‘Love’, ‘this’, ‘place’. We used various methods and libraries to tokenize such as NLTK, Gensim, Textblob. Tokenization can be performed on individual words or entire sentences. The process of breaking up text into words using separation technique is known



Figure 3.4: rating\_review, review\_len, word\_count, polarity.

as word tokenization, while the process of doing the same for sentences is known as sentence tokenization. We have used both word and sentence tokenization in our pre-processing. We have also used subword tokenization in which most often used words are assigned distinctive identifiers, and the less frequently used terms are divided into smaller words that best express the meaning on their own.

### 3.2.2 Stop Words Removal

This process is used to pull out the low-level information from the text data so that it could give more distinct value to the foremost information. It also helps to minimize the size of the dataset. With the help of stop words we can diminish the memory requirements while categorizing the reviews. The application of stop words are appealed in search systems, text classification applications, topic modeling, topic extraction and others[3]. An example of stop word removal is written below. Let us assume that all stop words are replaced with character such as P:

Original sentence = This is a food review of a restaurant

Sentence with stop words removal = P P P food review P P restaurant

### 3.2.3 Removal of rare words

This process is similar to the previous one as these types of words create more clustering and make the data size huge because of the high frequency so by removing these words our data become more precise and meaningful. It is important to remove rare words because in NLP it is also required to remove noisy characters such as



Figure 3.5: Visualization of text data using wordcloud.

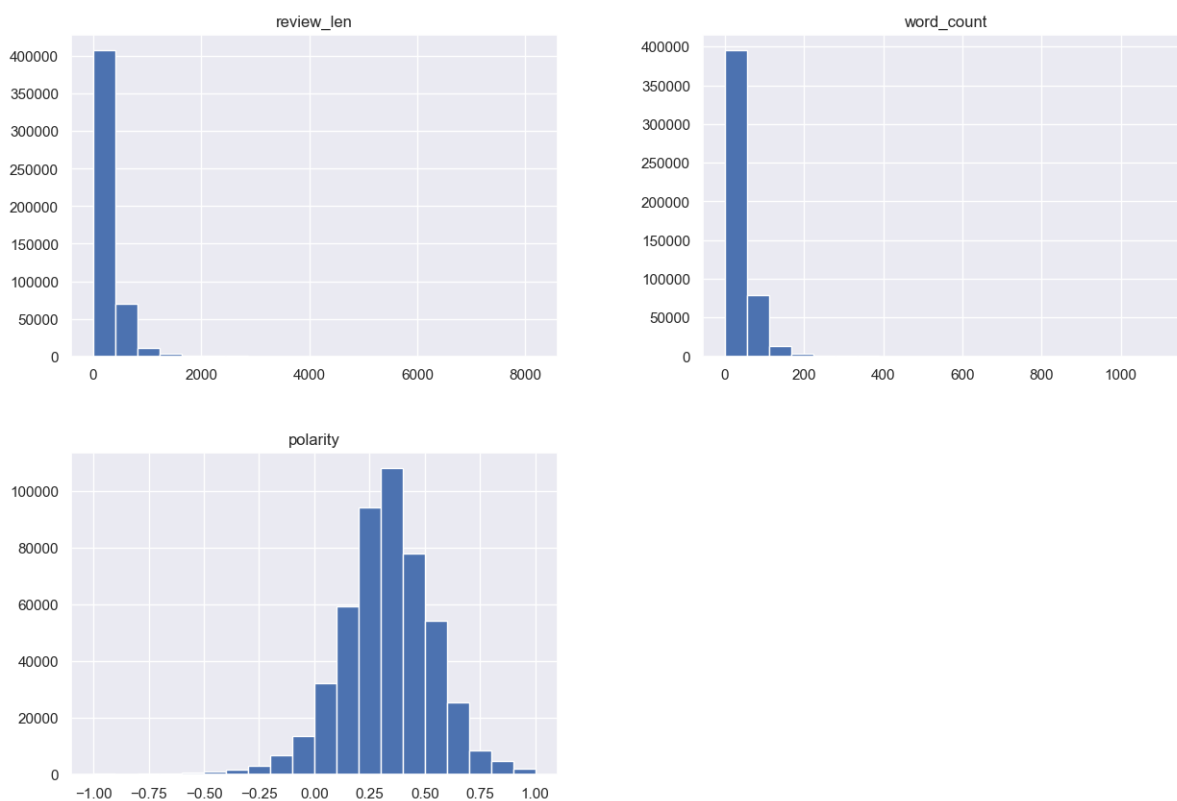


Figure 3.6: Histogram of review\_len, word\_count and polarity.

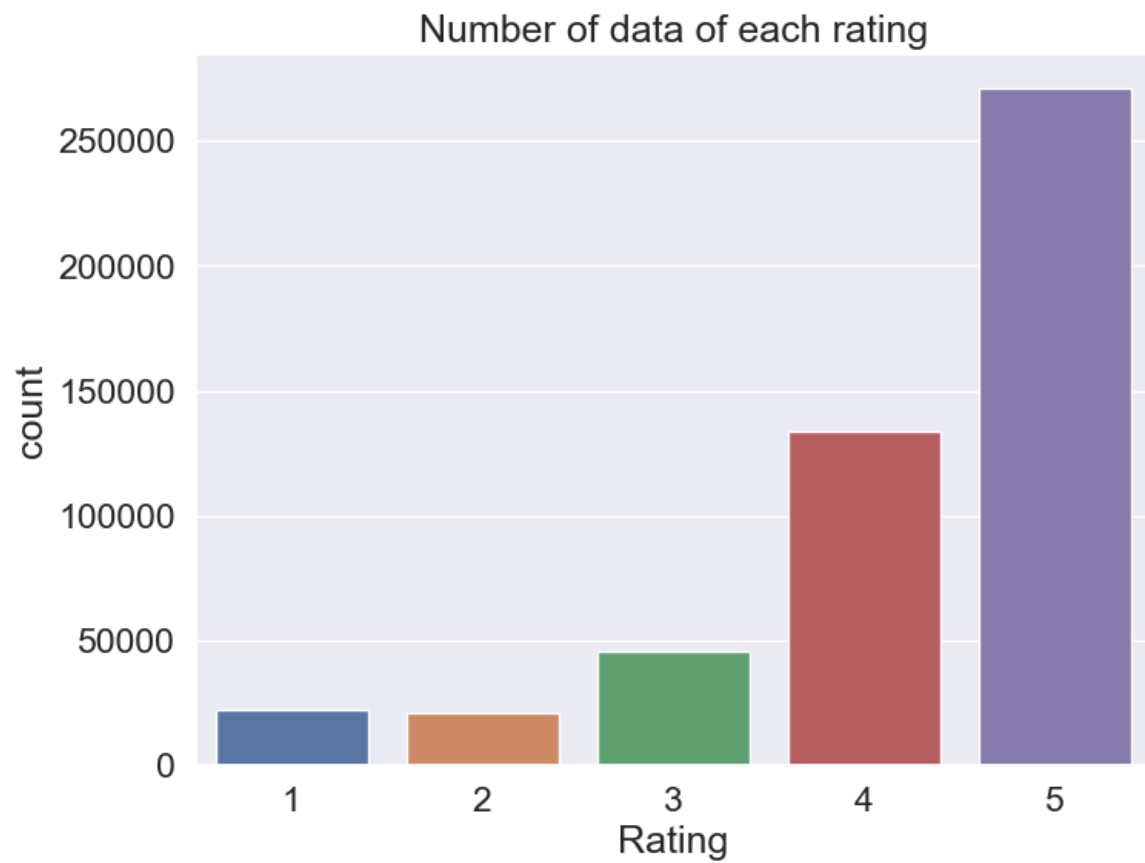


Figure 3.7: Number of data each rating.



Figure 3.8: Box plot of review rating of vs. polarity.





Figure 3.9: Point plot of product rating vs review length.

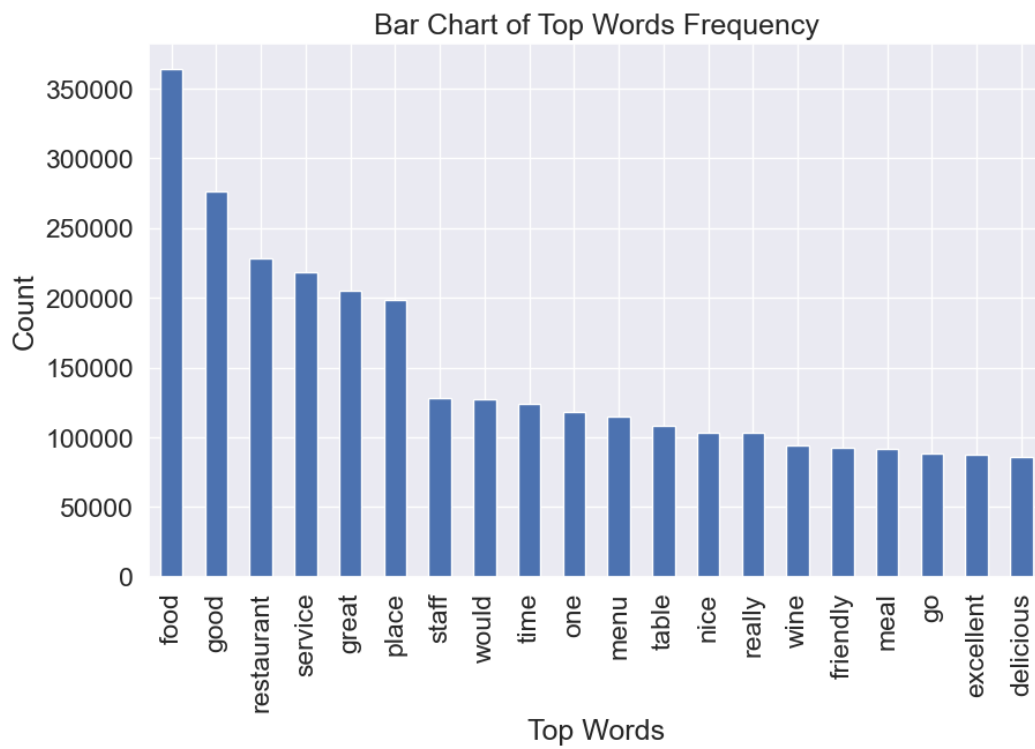


Figure 3.10: Bar chart of top world frequency.

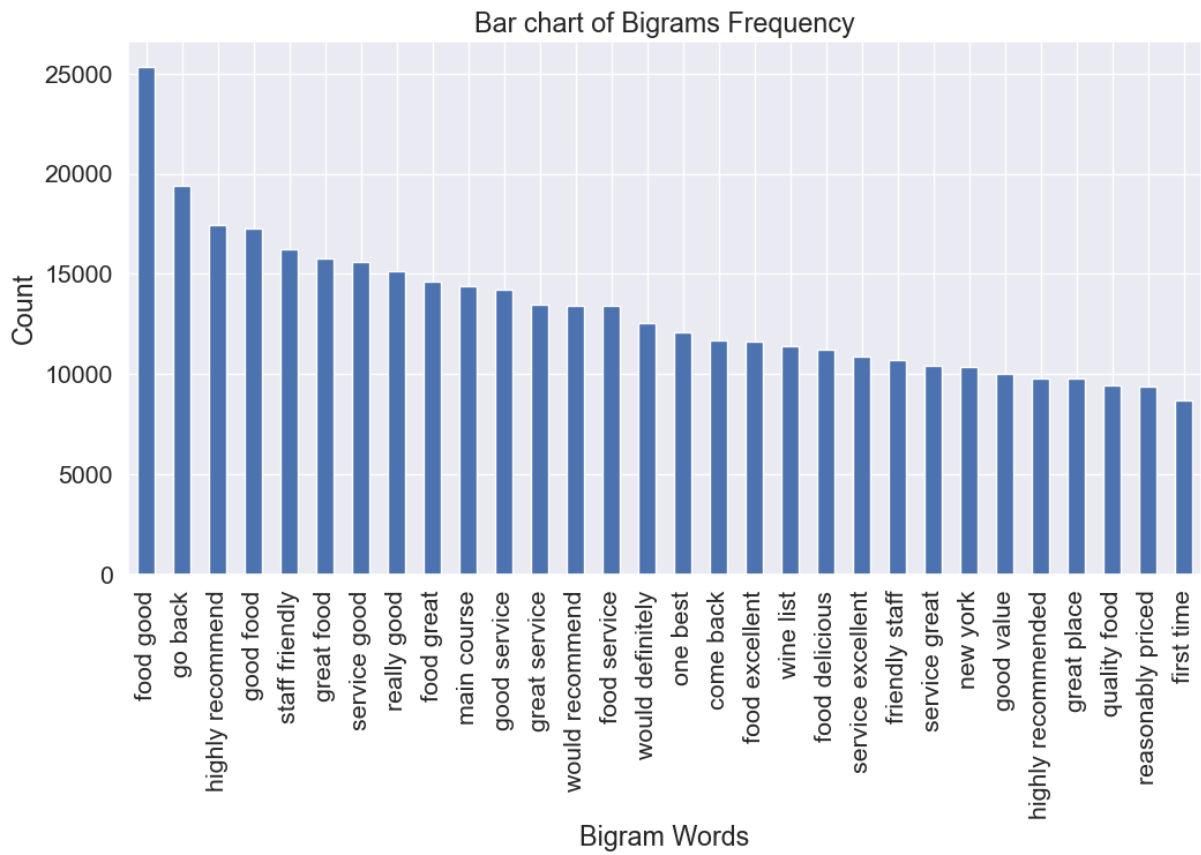


Figure 3.11: Bar chart of bigram frequency.

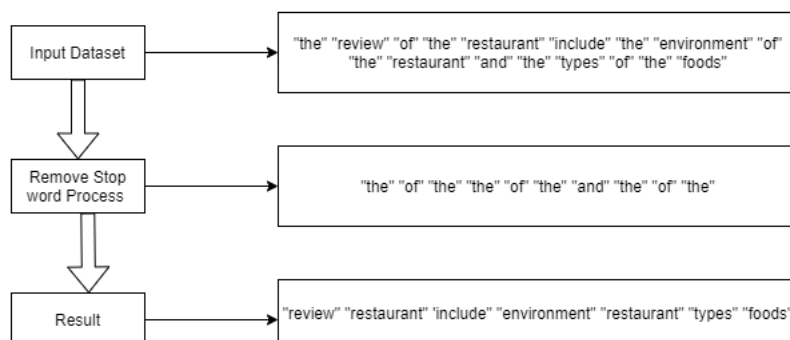


Figure 3.12: Stop words removal process.

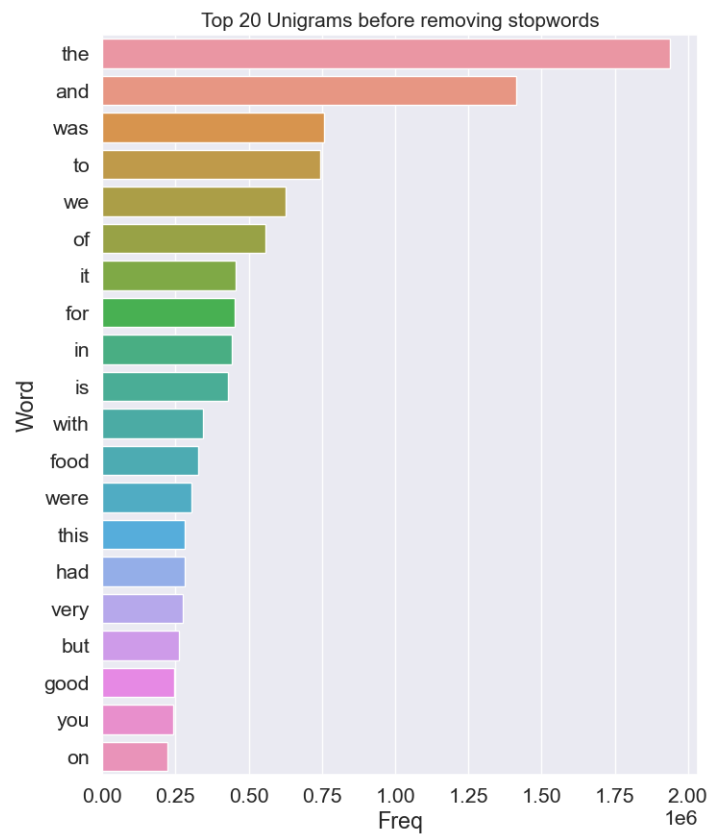


Figure 3.13: Top 20 unigrams before removing stopwords.

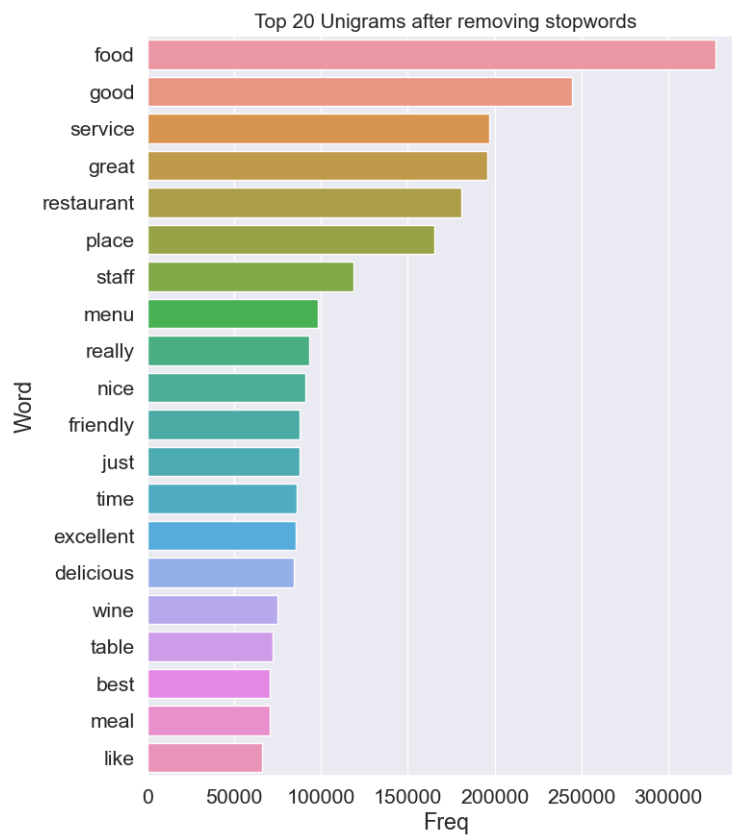


Figure 3.14: Top 20 unigrams after removing stopwords.

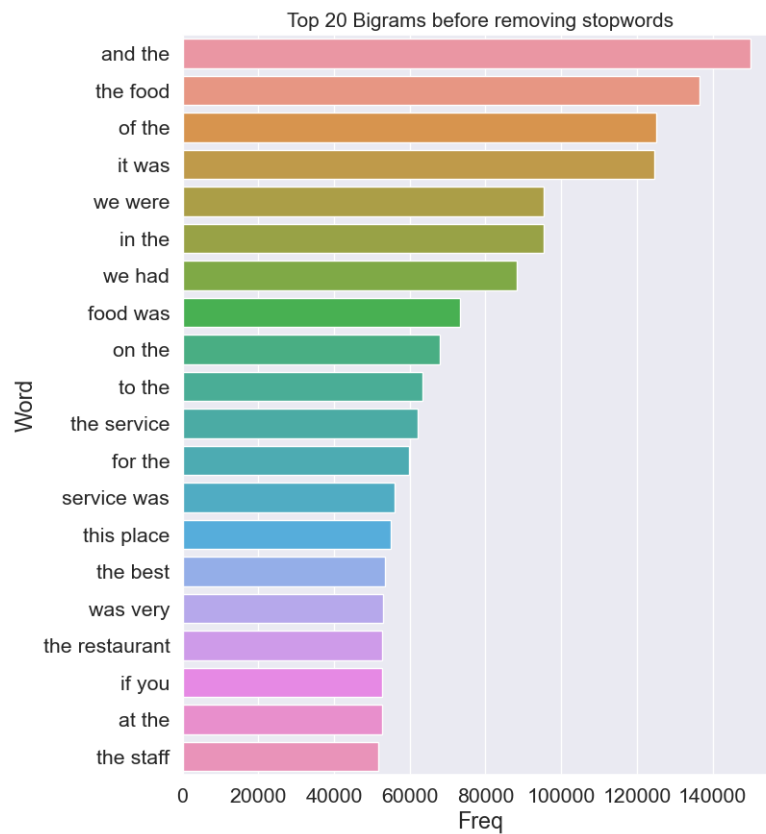


Figure 3.15: Top 20 bigrams before removing stopwords.

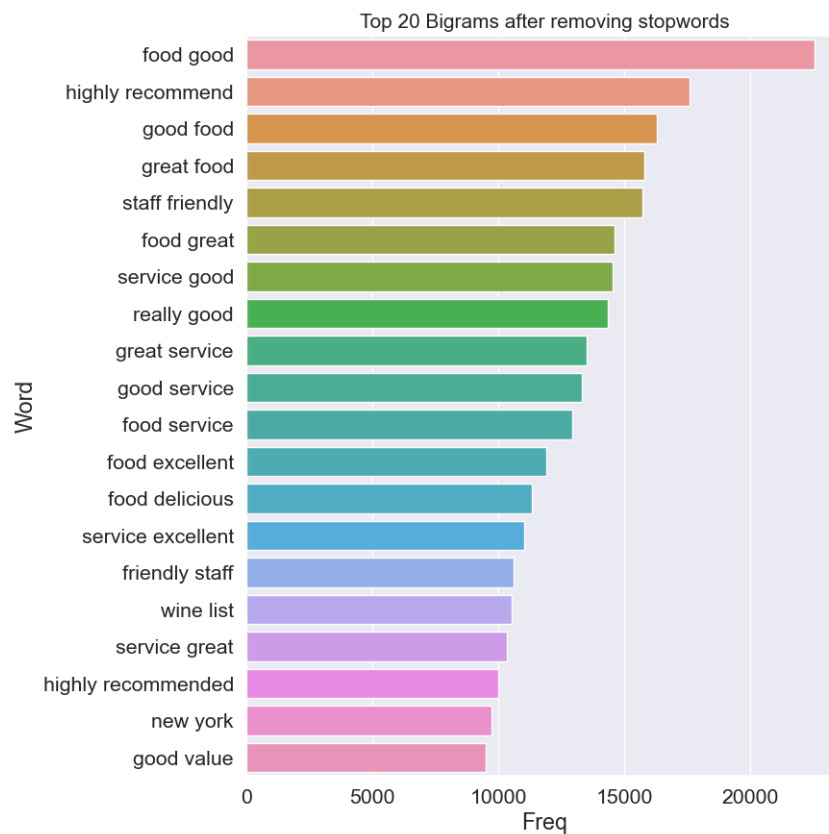


Figure 3.16: Top 20 bigrams after removing stopwords.

html leftouts and from words that are particularly unique in nature such as names, brands. We don't want these type of noisy tokens in our dataset.

### 3.2.4 Stemming

This is one of the techniques of Text Normalization that is used to prepare the words for further processing. The main purpose of doing this technique is so that it can normalize texts by removing prefixes and suffixes, such that different variations of a word are treated as the same word. There are some key factors that have impacts on stemming. First of all, obtaining through stemming represents the core meaning of a word known as stem. For example, the stem of the words "running," "runs," and "ran" is simply "run." Secondly, Stemming algorithms use linguistic rules and heuristics to identify and remove common prefixes and suffixes from words. By doing so, stemming aims to unify related words and reduce them to a common form. Besides, Stemming algorithms rely on predefined rules and patterns to perform the stemming process. These rules are designed to handle language-specific variations and morphological changes in words. Overall, stemming is a common technique used in NLP to normalize words and handle variations. It can be a useful preprocessing step for various text analysis tasks, although it may not capture the full semantic meaning of words compared to more advanced techniques like lemmatization.

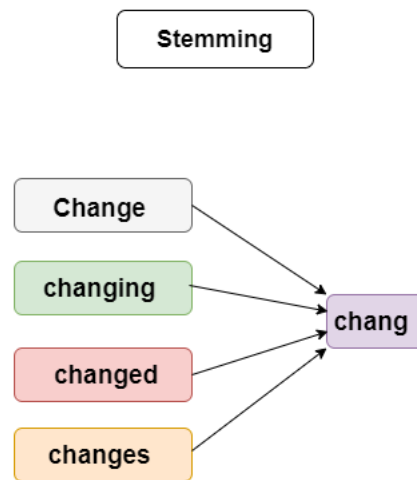


Figure 3.17: Process of Stemming.

### 3.2.5 Lemmatization

This is another technique for Text Normalization that is used to prepare the words for the process. This helps in reducing the dimensionality of text data and capturing the core meaning of words. There are some key points that have impacts on the process of lemmatization. First of all, The base or root form of a word obtained through lemmatization is the canonical form of the word, which represents its core meaning. For example, the lemma of the words "running," "runs," and "ran" is "run." Secondly, Lemmatization considers the grammatical views of words, such

as tense, gender, number, and part-of-speech tags, to determine the appropriate lemma known as morphological analysis. Moreover, Lemmatization is different from stemming, another text normalization technique. While stemming simply removes the suffixes or prefixes from words to obtain a root form, lemmatization takes into account the context and meaning of words. Overall, Lemmatization is a valuable technique in NLP for normalizing words and reducing the dimensionality of text data. It is commonly used in various text analysis applications to improve the accuracy and performance of machine learning models.

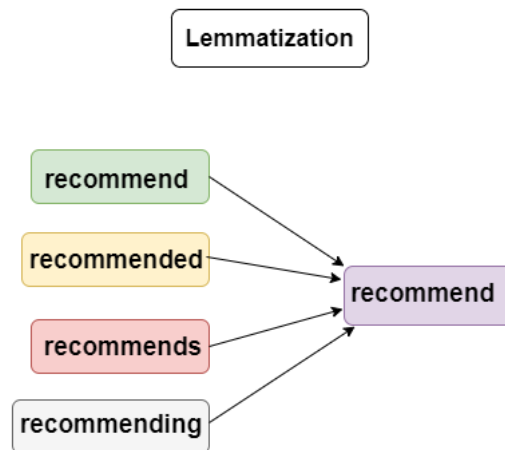


Figure 3.18: Process of Lemmatization.

### 3.2.6 Conversion of Emoji to Words

Likewise, emoticons, emoji can also represent important data which can be necessary for the dataset. Besides, emojis are connected with emotions which can be a good representative. Moreover, sometimes they represent more data than an actual text. So, we have converted emoji to words to keep the important data information into the dataset so that it would be an easy approach for any people to understand the emotions better and make decisions accordingly.

### 3.2.7 Removal of emojis

As emojis don't lay out any accommodating information that is why this process is used to remove those. It also helps to reduce the size of the data. Sometimes, they can give strong information about a text such as feeling expressions. Moreover, it can completely represent a different meaning if we remove the emoji.

For instance, if one person states a review about a restaurant in a sarcastic manner like this- "The food is out of the world (with a laughing face)" but the rating is very poor for that restaurant. With the help of the emoji, he presents his negative feedback in a sarcastic way. Besides, without emoji the whole text can represent a different meaning and can lead to a misunderstanding.

### 3.2.8 Removal of URLs

Individual URL removal is important as they tend to have such a low contingency individually. Uniform Resource Locators or URLs in a text point out as references to a location on the web. However, these do not deliver any subsidiary information. For that reason omitting URLs import a significant role.

### 3.2.9 Part of Speech Tagging

Part of speech tagging is a common natural language processing technique that involves classifying words in a text corpus in accordance with a certain component of speech. This includes nouns, verbs adjectives and other grammatical categories. Many NLP applications, such as information extraction, named entity identification, and machine translation, highly depends on POS tagging as it enable algorithms to comprehend a sentence's grammatical structure and to disambiguate words with numerous meanings, to improve the accuracy text classification and named entity recognition.

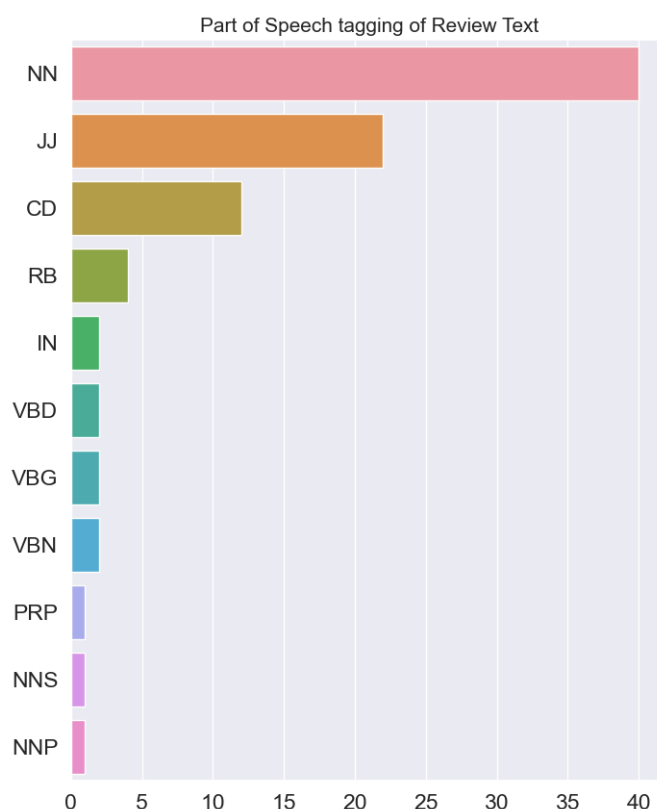


Figure 3.19: Part of speech tagging of review text.

## 3.3 Topic Modelling

Topic modelling is an unsupervised learning algorithm and is a technique which finds and assigns topics from a given corpus of present words. In the present time, all the data are in text and for that it is important to categorise the documents and

for that topic modelling is very important. For instance, in a restaurant, customers give reviews on different types of features. But it would be really hard to find all the features from the reviews. So to solve this problem, categorising the features from the reviews is very important. Topic modelling is a keyword that is extracted from the document and it is done based on the words present in the current document.

### 3.3.1 Latent Dirichlet Allocation

LDA is one of the techniques used for topic modelling. In LDA every word has a meaning. Firstly, latent means something is hidden and that is yet to be found. Secondly, Dirichlet designated that the model will make some assumptions on the topic of the document and the topics will be moving through a Dirichlet distribution. Thirdly, Allocation is to give away something from something unknown and in our part is the topic.

LDA is one of most used topic modelling which is a model of generative probabilistic and it is used for text corpora which is discrete data. This is formed of a hierarchical Bayesian model which is of three layers and each of the items is modelled as a composition of a finite amount of topics. Each of the topics is then modelled as an infinite composition for a collection of different topic probabilities, and then these topic probabilities return the exact representation of the context in the form of topic modelling. A word is defined as the base form of a discrete data which is an item or a token from a vocabulary indexed from 1..V. LDA assumes that the document are being generated by a generative statistical process, that each of the document is formed by the mixture of different topics and this topics are the mixture of different kinds of words. The figure shows 4 different topics which are distinct from one another and each topics has words which corresponds to the topics.

There are three types of hyperparameters. First one is  $\alpha$  ( ) which controls the number of topics in the documents. Second is  $\beta$  ( ) which holds the distribution of the number of words per topic in the document and third one is the  $k$  which is the number of topics to be extracted.

We must make certain inferences when picking an attribute since we lack labelled data. From the aforementioned evaluation using LDA, we deduce that the following are a few significant aspects that were cited in the majority of comments.

1. Food
2. Place
3. Service
4. Price

To get the proportions of these 20 topics for each review, we'll utilise the LDA model. This 20-vector will serve as our feature vector for supervised classification, with the aim of identifying whether a given statement is positive or negative.



```
[
(0,
'0.014*service" + 0.013*food" + 0.011*place" + 0.010*great" + '
'0.010*good" + 0.009*restaurant" + 0.009*nice" + 0.008*would" + '
'0.007*table" + 0.007*friendly"),
(1,
'0.017*good" + 0.016*food" + 0.012*u" + 0.012*place" + '
'0.011*restaurant" + 0.009*table" + 0.009*service" + 0.009*great" + '
'0.008*staff" + 0.006*time"),
(2,
'0.017*food" + 0.012*good" + 0.011*place" + 0.011*best" + 0.010*great" + '
'0.010*staff" + 0.010*service" + 0.009*go" + 0.007*u" + 0.007*table"),
(3,
'0.024*food" + 0.016*service" + 0.013*good" + 0.012*restaurant" + '
'0.012*would" + 0.010*great" + 0.010*recommend" + 0.010*u" + '
'0.009*staff" + 0.006*table"),
(4,
'0.016*good" + 0.012*service" + 0.009*food" + 0.007*great" + 0.007*one" + '
'0.007*meal" + 0.007*table" + 0.006*place" + 0.006*dinner" + '
'0.005*time"),
(5,
'0.022*good" + 0.016*food" + 0.014*place" + 0.012*great" + '
'0.008*service" + 0.008*staff" + 0.008*restaurant" + 0.007*time" + '
'0.007*dish" + 0.006*nice"),
(6,
'0.016*food" + 0.010*restaurant" + 0.009*place" + 0.009*;" + '
'0.008*staff" + 0.007*good" + 0.007*great" + 0.007*well" + 0.007*like" + '
'0.007*time"),
(7,
'0.018*food" + 0.018*restaurant" + 0.010*service" + 0.009*" + '
'0.009*place" + 0.008*good" + 0.008*would" + 0.007*great" + 0.007*one" + '
'0.007*u"),
(8,
'0.019*food" + 0.011*good" + 0.009*great" + 0.008*place" + '
'0.007*service" + 0.007*staff" + 0.006*would" + 0.006*time" + '
'0.006*restaurant" + 0.006*friendly"),
(9,
'0.017*good" + 0.016*food" + 0.010*restaurant" + 0.010*place" + '
'0.010*wine" + 0.010*great" + 0.010*service" + 0.008*staff" + '
'0.008*really" + 0.007*menu"),
(10,
'0.019*food" + 0.017*restaurant" + 0.013*good" + 0.012*service" + '
'0.010*place" + 0.009*great" + 0.007*time" + 0.007*staff" + 0.007*one" + '
'0.006*also"),

```

Figure 3.20: Keyword extracted using LDA (part1).

```
(11,
'0.023*food" + 0.015*service" + 0.012*great" + 0.012*good" + '
'0.012*place" + 0.012*restaurant" + 0.008*wine" + 0.007*excellent" + '
'0.007*time" + 0.006*would"),
(12,
'0.014*food" + 0.011*service" + 0.010*restaurant" + 0.009*great" + '
'0.008*place" + 0.007*good" + 0.006*one" + 0.006*table" + 0.006*would" + '
'0.006*dish"),
(13,
'0.018*food" + 0.015*great" + 0.013*restaurant" + 0.012*service" + '
'0.008*menu" + 0.008*good" + 0.007*wine" + 0.007*place" + 0.007*meal" + '
'0.006*u"),
(14,
'0.019*food" + 0.015*good" + 0.014*service" + 0.014*great" + '
'0.013*place" + 0.012*restaurant" + 0.008*time" + 0.007*bar" + 0.006*u" + '
'0.006*always"),
(15,
'0.016*food" + 0.011*good" + 0.011*restaurant" + 0.010*would" + '
'0.009*time" + 0.009*great" + 0.007*service" + 0.007*table" + 0.007*u" + '
'0.006*menu"),
(16,
'0.022*food" + 0.014*restaurant" + 0.014*great" + 0.010*service" + '
'0.009*staff" + 0.008*good" + 0.008*u" + 0.007*meal" + 0.007*dish" + '
'0.007*excellent"),
(17,
'0.021*food" + 0.018*great" + 0.012*good" + 0.011*service" + '
'0.009*place" + 0.008*delicious" + 0.007*restaurant" + 0.007*u" + '
'0.006*time" + 0.006*staff"),
(18,
'0.024*good" + 0.020*food" + 0.014*great" + 0.013*service" + '
'0.011*menu" + 0.010*restaurant" + 0.009*place" + 0.009*excellent" + '
'0.007*really" + 0.007*staff"),
(19,
'0.022*place" + 0.017*food" + 0.014*good" + 0.010*service" + 0.009*" + '
'0.008*restaurant" + 0.008*great" + 0.007*would" + 0.007*one" + '
'0.006*go")]
```

Figure 3.21: Keyword extracted using LDA (part2).

## 3.4 Classifications

### 3.4.1 Multinomial Naive Bayes

Multinomial Naive Bayes is a probabilistic machine learning algorithm that is the most widely used classification method in text mining [14]. It is basically based on the bayes theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. Besides, it is also dependent on some key factors. First of all, assumption is the most important key factor that represents the features are individually separate and distinct. In terms of document classification multinomial naive bayes is commonly used where features represent the frequency of the words. Secondly, feature representation is something in multinomial naive bayes that requires as the process goes on. It basically represents the requirement that the features need to be transformed into numerical values so that features can be represented by using a technique like the term frequency-inverse document frequency. Thirdly, in multinomial naïve bayes the word probability estimation is commonly used to calculate the probability of each class label given the feature values. Besides, the class with the highest probability is assigned as the predicted class label. On the other hand, to avoid getting zero probability multinomial naive bayes applies laplace smoothing which adds a small smoothing parameter (usually 1) to the feature counts and class counts to prevent zero probabilities. Overall, multinomial naive bayes is a simple and efficient algorithm for text classification tasks, especially when the independence assumption holds reasonably well. It is commonly used in spam detection, sentiment analysis, and document categorization. The distribution is parametrized by vectors  $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$  for each class  $y$ , where  $n$  is the number of features in text classification,  $\theta_{yi}$  is the probability  $P(X_i|y)$  of feature  $i$  appearing in a sample belonging to class  $y$  [16]. The formula is given below:

$$\theta_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad (3.1)$$

### 3.4.2 Random Forest

A flexible ensemble machine learning technique called random forest is frequently employed for classification and regression problem-solving. Basically, it’s a collection of decision trees, where each and every tree is made of a random subset of features and the final prediction is a combination of all the predictions of individual trees. Random forest has some special key features that are required to be noticed. First of all, assemble each individual decision tree and combine them to make a prediction known as assemble of decision tree in random forest algorithm. Besides, random forest algorithms train each decision tree with different random data and features. Secondly, random forest introduces randomness by considering only a subset of features at each split during the construction of decision trees. This random feature selection helps to decorrelate the trees and increase diversity, leading to better overall predictions. Moreover, random forest use an algorithm called CART to construct any decision tree and the trees are grown until a stopping criterion is met, such as reaching the maximum depth or having a minimum number of samples per leaf. For classification tasks, Random Forest combines the predictions of individual trees

using majority voting. The class with the most votes across all trees is assigned as the predicted class label. For regression tasks, the predictions of individual trees are averaged to obtain the final prediction. Furthermore, there are few other key factors such as Out-of-Bag (OOB) estimation, Model training, Feature importance has an impact on random forest algorithm.

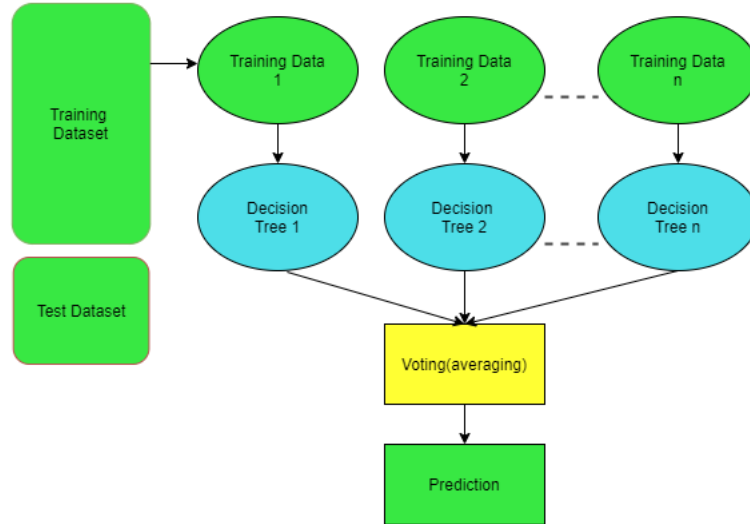


Figure 3.22: Random forest algorithm.

### 3.4.3 Support Vector Machine

Both classification and regression applications use the efficient supervised machine learning technique Support Vector Machines (SVM). In order to categorize the data points into groups or predict a continuous objective variable, SVM looks for the optimum hyperplane. This algorithm is influenced by a few significant elements [16]. First, in SVM, the data points are divided into several classes using a hyperplane. Additionally, the primary goal of SVM is to identify the hyperplane that maximizes margin, which is essentially the distance between the hyperplane and the closest data point for each class. Second, SVM handles nonlinear decision boundaries by using kernel methods[8]. SVM may implicitly transfer the initial input space into a higher-dimensional feature space thanks to that kernel method where the data points may be divided using a linear hyperplane. Thirdly, the data points nearest to the hyperplane—called support vectors—determine the hyperplane’s location and orientation. These support vectors are essential to SVM since altering their placements can have an impact on the decision boundary and the hyperplane. Fourth, by adding a slack variable, SVM can handle situations when the data points cannot be separated linearly. Soft margin classification strikes a compromise between maximizing the margin and tolerating some mistakes by allowing certain data points to be misclassified or fall within the margin. Additionally, the regularization parameter ( $C$ ) in SVM regulates the trade-off between maximising the margin and reducing classification mistakes. Wider margins are produced by a lower  $C$  value. Both classification and regression applications use the efficient supervised machine

learning technique Support Vector Machines (SVM). In order to categorize the data points into groups or predict a continuous objective variable, SVM looks for the optimum hyperplane. This algorithm is influenced by a few significant elements.. First, in SVM, the data points are divided into several classes using a hyperplane. Additionally, the primary goal of SVM is to identify the hyperplane that maximizes margin, which is essentially the distance between the hyperplane and the closest data point for each class. Second, SVM handles nonlinear decision boundaries by using kernel methods. SVM may implicitly transfer the initial input space into a higher-dimensional feature space thanks to that kernel method where the data points may be divided using a linear hyperplane. Thirdly, the data points nearest to the hyperplane—called support vectors—determine the hyperplane’s location and orientation. These support vectors are essential to SVM since altering their placements can have an impact on the decision boundary and the hyperplane. Fourth, by adding a slack variable, SVM can handle situations when the data points cannot be separated linearly. Soft margin classification strikes a compromise between maximizing the margin and tolerating some mistakes by allowing certain data points to be misclassified or fall within the margin. Additionally, the regularization parameter (C) in SVM regulates the trade-off between maximising the margin and reducing classification mistakes. Wider margins are produced by a lower C value.

Furthermore, model training has a larger impact on SVM algorithms too. Overall, Support Vector Machines are widely used in various applications, including text categorization, image classification, and bioinformatics. They are known for their ability to handle high-dimensional data, provide robust generalization, and handle both linear and nonlinear decision boundaries through the use of the kernel trick.

# Chapter 4

## Result & Analysis

Different kinds of topics that were found were converted to feature vectors. Afterward the feature vector is separated in 80% and 20% training and testing. Then we have converted our rating into positive and negative which is denoted as 1 and 0 .We have implemented three linear classification algorithms and made five small samples of training set using k-fold. Then we calculated the accuracy of each of the samples and finally we found the accuracy of the models. Furthermore we have found the precision , recall and the f1 score of each of the models.

Table 4.1: Result analysis

Algorithm	Accuracy	Precision	Recall	F1 Score
SVM	0.911551	0.91155	0.911551	0.911551
MNB	0.903327	0.903327	0.903327	0.903327
RandomForest	0.900671	0.900671	0.900671	0.900671

The estimated analysis is shown as a confusion matrix. It illustrates the number of accurate and wrong estimates made for each class. It aids in clarifying the classes that model mistakes for other classes. The confusion matrix was used to assess how well the procedures employed after the categorizations were performed. A confusion matrix is used to demonstrate the performance of a classification system. The results of a classification algorithm are shown and summarized in a confusion matrix. The layout of the binary classification confusion matrix is shown in Fig. 4.1.

Figure 4.1: Confusion Matrix.

### 4.0.1 Confusion Matrix for Naive Bayes

For naive bayes , for target 0 it cannot classify target 0 at all but can completely classify target 1. Furthermore, for target 0 it has more false positives which means that the model cannot classify the target 0.

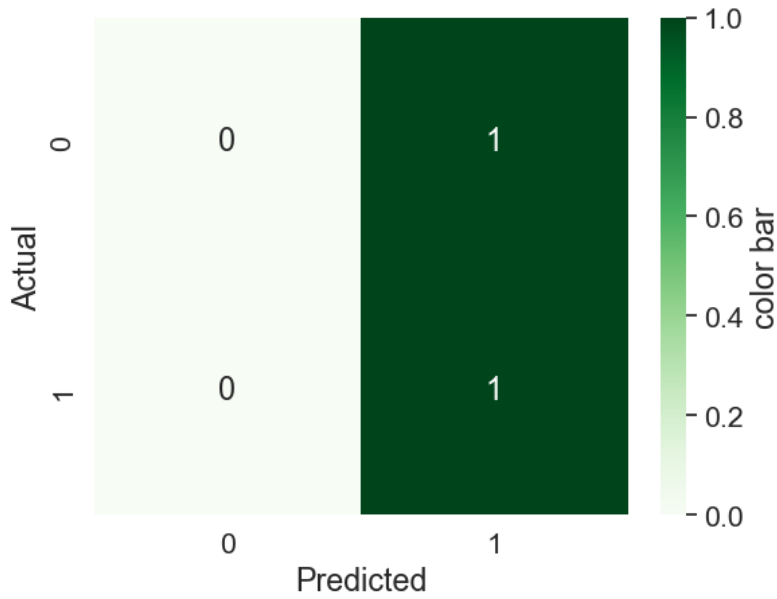


Figure 4.2: Confusion Matrix for Naive Bayes

#### 4.0.2 Confusion Matrix for SVM

From the confusion matrix of SVM it is visible that the model has true positive for target 1 but for target 0 the model does not have good true positive instead it has more false positive and false negative.

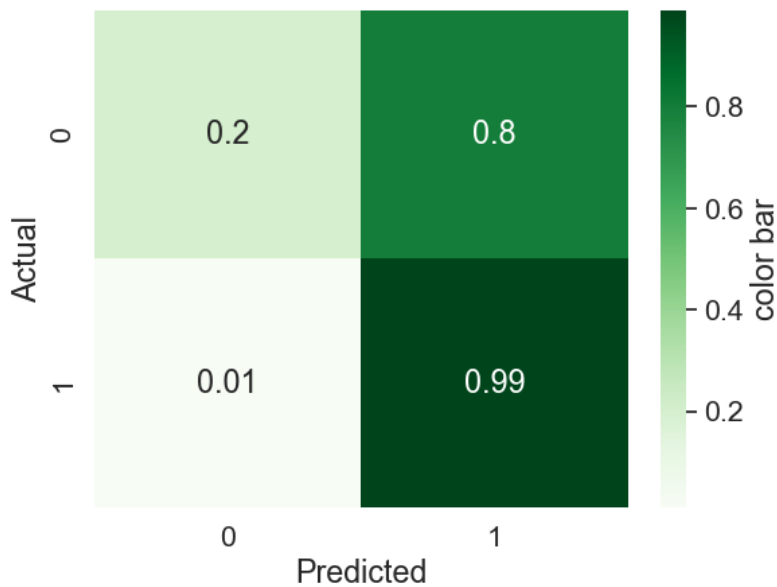


Figure 4.3: Confusion Matrix for SVM

#### 4.0.3 Confusion Matrix for Random Forest

Similarly , for random forest , the model shows true positive higher for target 1 compared to target 0.

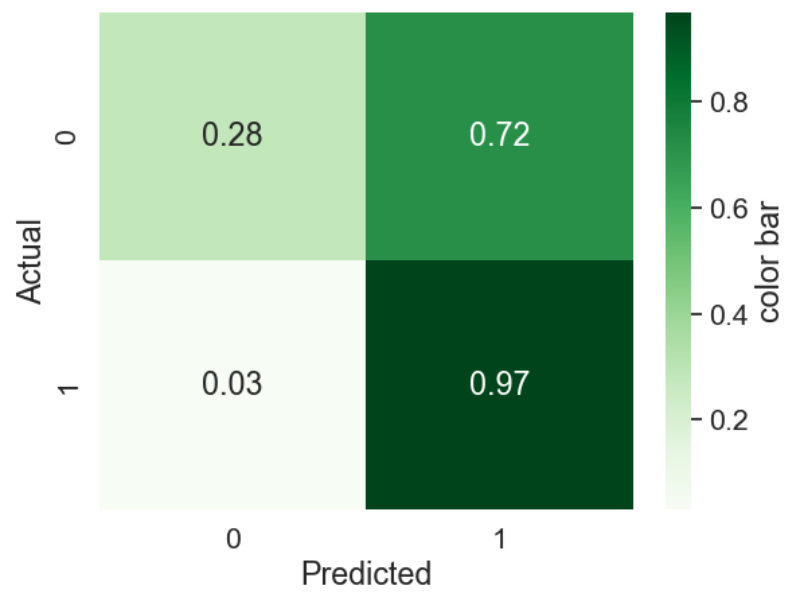


Figure 4.4: Confusion Matrix for Random Forest

## Chapter 5

### Conclusion

Data mining, computational linguistics, and natural language processing are all parts of opinion mining. Based on the review, it decides if the consumer has a favorable or unfavorable impression. Large amounts of feedback from customers about any restaurant may be handled using the technology, which also offers improved legitimacy. This article focuses on putting into practice an aspect-based opinion miner in consumer domains like restaurant reviews, which automatically identifies significant aspects and opinions of a restaurant through reviewing reviews, then creates a sentiment profile of each restaurant, which can then be used to compare and choose restaurants in a specific location by any customer. There may be room for improvement in this essay. The same, as well as an examination of other varieties, is what our next study seeks to offer. In order to increase the precision of opinion mining, our further study seeks to incorporate the aforementioned as well as the analysis of various sentence types, such as comparative and conditional sentences. For improved outcomes, classifiers can also be used to examine the suggested system.



# Bibliography

- [1] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, “Guest editorial special issue on concept-level opinion and sentiment analysis,” *IEEE Intelligent Systems Magazine, Special Issue on Concept-Level Opinion and Sentiment Analysis*, vol. 28, pp–15, 2012.
- [2] B. Liu, “Sentiment analysis and opinion mining,” *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- [3] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, and I. Mathur, *Natural language processing: python and NLTK*. Packt Publishing Ltd, 2016.
- [4] S. Poria, E. Cambria, and A. Gelbukh, “Aspect extraction for opinion mining with a deep convolutional neural network,” *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016.
- [5] S. Poria, E. Cambria, and A. Gelbukh, “Aspect extraction for opinion mining with a deep convolutional neural network,” *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016.
- [6] S. Ruder, P. Ghaffari, and J. G. Breslin, “A hierarchical model of reviews for aspect-based sentiment analysis,” *arXiv preprint arXiv:1609.02745*, 2016.
- [7] K. Bauman, B. Liu, and A. Tuzhilin, “Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 717–725.
- [8] A. Navlani, *Scikit-learn svm tutorial with python (support vector machines)*, Dec. 2019. [Online]. Available: <https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python>.
- [9] C. Gong, J. Yu, and R. Xia, “Unified feature and instance based domain adaptation for aspect-based sentiment analysis,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7035–7045.
- [10] B. D. P. Statista, “Number of social network users worldwide from 2017 to 2025,” URL: <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/> : 06.02. 2022), 2020.
- [11] H. Xu, B. Liu, L. Shu, and P. S. Yu, “Dombert: Domain-oriented language model for aspect-based sentiment analysis,” *arXiv preprint arXiv:2004.13816*, 2020.
- [12] S. Sazzed, “A hybrid approach of opinion mining and comparative linguistic analysis of restaurant reviews,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021, pp. 1281–1288.

- [13] I. L.-R. Botana, V. Bolón-Canedo, B. Guijarro-Berdiñas, and A. Alonso-Betanzos, *Explain and conquer: Personalised text-based reviews to achieve transparency*, 2022. arXiv: 2205.01759 [cs.LG].
- [14] A. A. Awan and A. Navlani, *Naive bayes classifier tutorial: With python scikit-learn*, Mar. 2023. [Online]. Available: <https://www.datacamp.com/tutorial/naive-bayes-scikit-learn>.
- [15] [Online]. Available: <https://datareportal.com/social-media-users%5C%7D,%20journal=%7BDataReportal%7D>.
- [16] [Online]. Available: <https://scikit-learn.org/stable/modules/svm.html>.