TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PULCHOWK CAMPUS

THESIS NO: M-123-MSTIM-2020-2022

Sentiment Analysis of Different E-commerce platform reviews using Machine Learning Algorithm

by
Manil Vaidhya

A THESIS
SUBMITTED TO DEPARTMENT OF MECHANICAL AND AEROSPACE ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN ENGINEERING IN TECHNOLOGY AND INNOVATION MANAGEMENT

DEPARTMENT OF MECHANICAL AND AEROSPACE ENGINEERING
LALITPUR, NEPAL

SEPTEMBER, 2022

COPYRIGHT

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

DEPARTMENT OF MECHANICAL AND AEROSPACE ENGINEERING

The undersigned certify that they have read, and recommended to the Institute of Engineering for acceptance, a thesis entitled "**Sentiment Analysis of E-commerce Reviews using Machine Learning Algorithm**" submitted by Manil Vaidhya in partial fulfillment of the requirements for the degree of Master of Science in Engineering in Technology and Innovation Management.

_____

Supervisor, Assistant Professor Sanjaya Neupane
Department of Mechanical and Aerospace Engineering

_____

Supervisor, Assistant Professor Aayush Bhattarai
Department of Mechanical and Aerospace Engineering

_____

External Examiner, Senior Lecturer Dipesh Shrestha
Kantipur Engineering College

_____

Committee Chairperson, Associate Professor,
Dr. Surya Prasad Adhikari
Head of Department
Department of Mechanical and Aerospace Engineering

Date: September 25, 2022

ACKNOWLEDGMENT

I would like to express my deepest sense of gratitude and thanks to my supervisors Assistant Professor Sanjaya Neupane and Assistant Professor Aayush Bhattarai, Department of Mechanical and Aerospace Engineering, Pulchowk Campus for their invaluable guidance and encouragement. Their useful suggestions and continuous motivation are sincerely acknowledged. My special thanks goes to Dr. Sanjeev Maharjan, Program Coordinator of Master in Technology and Innovation Management for providing me guidance on how to choose a research topic and how to work on a thesis.

I would like to express my sincere thanks to Associate Professor Dr. Surya Prasad Adhikari, HOD of Department of Mechanical and Aerospace Engineering, for giving me this opportunity to undertake this thesis. Also, I would like to thank all my teachers and colleagues for their direct and indirect help throughout the thesis

ABSTRACT

E-commerce provides different products/services to the customer where customers can easily get their desired products anywhere they want. While buying online, they rely on the product reviews made by other users which gives much more emphasis on the product review as it is required for the selection of a product. For the analysis of such reviews, sentiment analysis is done. Since the data are in huge numbers, machine learning algorithms are used for the fast and effective calculation and analysis of these product reviews. These reviews can be done quickly using machine learning as the model is created where thousands of reviews are made. For the better accuracy of the model, the datasets are subjected to different pre-processing techniques. Then, both supervised and unsupervised learning methods as well as the deep learning method to classify the sentiments of the dataset in positive or negative class. For the validation of our model, secondary dataset were obtained from the ecommerce platforms like Daraz. For supervised learning models, we have used Naive Bayes and SVM classifiers. From lexicon based analysis, VADER classifier is used which exhibits 68% accuracy when validating with the secondary data. Also supervised algorithms like SVM classifiers exhibit 71% accuracy whereas Naive-Bayes classifiers exhibit 68% accuracy for the data gathered. But the highest accuracy was obtained from deep learning models which exhibit 75% accuracy. Thus, the classification of datasets were done into two classes positive and negative and the best result was obtained for Daraz among different e-commerce sectors as the no. of positive sentiments are much higher than negative sentiments.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BOW	Bag of Words
CNN	Convolutional Neural Network
ENN	Elman Neural Network
IDF	Inverse Document Frequency
ML	Machine Learning
NB	Naive Bayes
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PCA	Principal Component Analysis
POS	Parts of Speech
RMSE	Root Mean Square Error
SA	Sentiment Analysis
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term frequency-inverse document frequency
UI	User Interface
URL	Uniform Resource Locator
UX	User Experience

CHAPTER ONE: INTRODUCTION

1.1 Background

As we all know, the success of a company or a product directly depends on its customers. So, if the customer likes your product, it's your success but if they don't like it, there needs to be some improvement with the product or delivery by analyzing the reviews made by the customers which can be done through analyzing the sentiments of the customers reviews. Thus sentiment analysis is an important field of analysis where we can get the approximate sentiments of the customer regarding a certain product or services of any company. Sentiment analysis can be used for determining whether the data is either positive, negative or neutral for monitoring the business brand and understanding what the customer needs.

Online reviews are a valuable resource for consumers choosing what to buy, what to watch, where to eat, and how to spend their money (Pandey & Soni, 2019). These sources offer evaluations and comments on both their services and products, allowing viewers to rate. Since the beginning of the 2000s, SA has been active. SA falls under the umbrella of web content mining. Sentimental analysis (SA) creates a clean dataset that will be utilized for feature extraction by utilizing a variety of pre-processing techniques such as tokenization, vectorization, and the removal of duplicate data. Different techniques are available to delete those unnecessary data. Data can be extracted directly from the e-commerce organization or market survey can be done with some questionnaires and analysis must be done in the proper way for the best possible sentiment outcome of that particular text, sentences. SA can be used in market intelligence to gauge consumer happiness with goods and services, identify areas for improvement, forecast pricing changes, and do a host of other tasks that help companies create new goods and services while also enhancing existing ones in response to consumer feedback.

We occasionally encounter noisy text data, which contains spelling errors, idioms, sarcasm, and informal words and phrases that are not recognized by the SA's present methodologies. As a result, Cluster Computing automatically analyzes such data, identifies the appropriate emotion contained in text, and facilitates the decision-making process for consumers. Almost all successful businesses nowadays must use sentiment analysis to aid with customer satisfaction requirements. Deep Learning (DL) techniques, a branch of machine

learning that is mostly employed in SA systems, are now being researched by a number of academics to enhance the processes involved in manipulating data. The most accurate Sentiment Classification for SA of online product reviews is achieved using LSTM, LSIBA-ENN, etc., which operate on the RNN concept. Furthermore, many new preprocessing techniques are developed for the detailed analysis of those sentiments. As sentiments can be tricky to classify because of the mood swings or different external factors, proper preprocessing techniques must be introduced to classify those datas. Classification of those preprocessed data are used to classify whether they belong to positive sentiments or negative sentiments. Classification depends on the nature of the datasets. If the datasets are unsupervised, unsupervised algorithms like lexicon analysis, textblob etc are used to classify the sentiments. Similarly, if the datasets are supervised, supervised algorithms like Naive Bayes, SVM, KNN etc are used to classify the sentiments. As for the larger dataset where computation takes a lot of time, deep learning methods are commonly used to classify the sentiments (Vaidhya et al., 2017).

People mostly express their views/opinions on the discussion forms like comment section, review section and any other platforms. These reviews are generally used for the development of an organization as in to improve or make changes along with what the users expect with the product or services or their opinions. The amount of data creates a problem to carefully analyze the exact sentiment of the people and therefore sentiment analysis comes in handy as the user's text and sentiment towards an organization or their product or services can be analyzed through the positive and negative sentiment obtained through the sentiment analysis. The main task in this case is to use different machine learning algorithms in order to make the computation and classification of the classes much faster and accurate. We can maintain the accuracy and different evaluation techniques to provide the consistent and more accurate results with the help of machine learning algorithms as they are devised to perform such tasks with much ease and efficiency (Wang, 2017). Hybrid sentiment analysis models are the most modern, efficient, and widely-used approach for sentiment analysis. Provided you have well-designed hybrid systems, you can actually get the benefits of both automatic and rule-based systems. Hybrid models can offer the power of machine learning coupled with the flexibility of customization. Also, Two techniques of neural networks are common – CNN or Convolutional Neural Networks for processing of images and RNN or Recurrent Neural Networks for NLP tasks.

1.2 Problem Statement

Many business related industries are trying hard to provide a quality service and maintain their stability in the market. Not many people try to take in people's opinion resulting in the instability in the market. All the market products are consumed by the people in a specific range. For example: Customers aged from 15-30 are much more interested in electronics gadgets, fashion attire, entertainment etc. While the people from age group 30 and above are interested in health related products etc. For this very reason we have to understand the people's needs and their expectations with the product. This is a very serious problem in this pandemic era as more and more people are using the ecommerce site for selection of the products. Some are interested in quality products, some are interested in cheap products, some are interested in both. Maintaining such requires an understanding of the customer's desires. In many ecommerce sites, there is a review section where users explain their feelings towards a product.

So to avoid that and understand the customer's need and their behavior about the product, sentiment analysis can be done for such reviews to understand the customer's point of view about a product. This will have a direct impact on the organization selling the product as what needs to be done so that customers will be engaging on their site for similar or different kinds of product. Due to which maintaining an engaging relationship with customers is very much important. Customer satisfaction towards a product holds the key to success for every organization. So the product data must be carefully analyzed whether the expressed opinion by the customer is either positive, negative or neutral.

1.3 Research Objective

1.3.1 Main Objective

The main objective of this thesis work is to find the sentiment of different Ecommerce platforms reviews using Machine Learning Algorithm.

1.3.2 Specific Objectives

The specific objectives of this research are as follows:

- To find the class of sentiments (Positive or Negative) using different algorithms

- To perform analysis for different e-commerce platforms based on their number of positive and negative sentiments reviewed by the users.

1.4 Limitation

The sentiment of a particular text cannot be 100% accurate while classifying using different algorithms and mathematical and logical analysis. Certain texts like sarcasm, idioms etc. cannot reflect the true meaning of the sentiment while using certain algorithmic techniques. Besides, images like emojis or emoticons also reflect some sentiment which require a deep level of pre-processing. While the research is based on the data available from the questionnaires created, some important dependent factor always goes missing while finding the sentiments of the users. Also, the main weakness of this procedure is that the researcher performs carelessly or does not follow systematic procedures. Besides above, the following limitations have been observed.

- Data may contain sparse values and huge amounts of features so appropriate algorithms must be used for pre-processing.
- Sentiment analysis of review data are often much less accurate when extracted from other domains such as news or social media because of the differences in how people express themselves in these domains
- Sentiments are very likely to change over time in accordance with a person's mood.

CHAPTER TWO: LITERATURE REVIEW

2.1 Relevant Theories

Some details of the theoretical background that involves sentiment analysis. As we know that finding the correct sentiment of a particular text, sentences or opinions is the main task of sentiment analysis. To make it happen the language that we use to communicate must be broken into computer/machine level language where a particular word has a weighted value for a particular sentence/text towards positive or negative sentiments.

2.1.1 Natural Language Processing

The field of NLP has outgrown it in the coming year. Initially confined to gathering data from a limited set of digitized documents, the advent of the World Wide Web saw an explosion in information in many different languages. One of the main reasons to use NLP techniques is to use it for applications like speech recognition, text analytics etc. Following are the key approaches in NLP

2.1.2 Preprocessing/Tokenization

The first task is to segment a given document into words or sentences. Almost all languages use white space as the delimiter but this can create some issues as some words may contain a whitespace and they are whole as a word, so when there is the use of whitespace a separate pre-processing must be done. Some of the examples of such words that include white space between words but are the same words are “I’m” into “I am” and deciding whether or not to separate a word such as “well-known” into two words.

2.1.3 Lexicon Analysis

After processing the text, the next task is to divide the text into separate words. A word or lexeme in linguistics represents a meaning and this meaning is different for different languages. Lexemes which are mostly composed of two types of morphemes - bound and free. The bound morphemes are words that have suffixes and affixes like “tion” and “un”. While the free morphemes are just independent words without the attachment of such affixes and suffixes like “dog”.

2.1.4 Semantic Analysis

Till now all the above preprocessing is done but the question still remains whether those pre-processed words hold any meaning or not. Let's consider the following sentence:

I am going down
I am feeling down
I am walking down

The above three sentences have different meanings like the first sentence tells that he is going down the stairs or floors or he is feeling ill. Similarly, the second sentence can have a meaning that he is ill. The third sentence can have a meaning that he is walking somewhere else. These are the traditional approaches to process a particular text or sentence by doing some preprocessing like tokenization. However from the above examples we can see that as the computational and processing task increases, the complexity to process the output also increases. That is why the ML approach is much popular nowadays as it negates the difficulties in the traditional approach and using the statistical or ML approach will make the result much better. With the increase in such complexities, the ML approach proposes two learning methods for NLP classification.

1. Supervised Learning Method
2. Unsupervised Learning Method

2.2 Related Works

A large amount of research based on sentiment is often carried out on the user's reviews about a particular product or service using different techniques that involve minimal to maximum number of accurate results.

Some of the related studies are done by Zhao, H. (2021) performed a sentiment classification by using LSIBA-ENN (Local Search Improvised Bat Algorithm based Elman Neural Network) algorithm that takes input from the feature vector which are preprocessed by different techniques like tokenization, lemmatization and stemming. The product reviews were gathered by scraping the website and processed for feature extraction (Zhao et al., 2021). The proposed LSIBA-ENN's performance is analyzed with the existing SVM (Support vector machine), NB(Naive Bayes), and ENN techniques regarding '4'

performance metrics: precision, recall, f-measure, and also accuracy. The accuracy was high for the SVM model following ENN and NB. How-ever this research utilized the dataset of apps and movies and the use of deep learning methods could help this research evaluate the data of multimedia data sets with much more precision.

Similarly, Raj P M, K. & Sai D, J.(2021) used a VADER sentiment analysis algorithm to classify the sentiment and used LDA for pre-processing the data inputs (P M & D, 2021). The Vader social sentiment analysis algorithm performed 15% better in overall accuracy. But, the classification was difficult as the data increased which may lead to negative approaches or False-Negative parts resulting in the decrease of accuracy. Vader classification is done for the unsupervised dataset. The author obtained the unsupervised dataset and used the VADER classification techniques that mostly uses rule based analysis. This common method is also known as lexicon analysis as certain rules are followed for the classification of the sentiments.

Another researcher E. Suganya & S. Vijayarani (2020) proposed a scraping technique to collect the data from multiple sites like amazon, flipkart, snapdeals etc. to be used for sentiment analysis using machine learning algorithms like KNN, SVM, Random Forest, CNN and Hybrid SVM-CNN (Suganya & Vijayrani, 2019). Out of these Hybrid SVM-CNN had the highest accuracy in case of performance. Generally Supervised learning produces a great model which has the efficient accuracy but when applying the deep learning method, the accuracy tends to increase as the model is subjected to various hidden layers and back propagation allows the data to be much more efficient resulting in the increase in accuracy.

Meanwhile C.Zucco, B.Calabrese (2018) designed a model for sentiment analysis of so-cial network and text mining. Using the BOW (Bag-of-Words) and N-grams method for feature extraction, and again the output is subjected towards SVM model and lexicon model for classification (Zucco et al., 2019). The Lexicon model has a higher accuracy than the SVM model. But SVM has a higher precision than the Lexicon model. While the accuracy of the SVM model is always higher, it all depends on the data and the pre-processing of those data. Pre-processing of those data actually plays a bigger role in increase of accuracy. So if the pre-processing of the data is done effectively, the Lexicon method can have a higher accuracy.

Esha Tyagi (2017) proposed a SVM classifier for different datasets. These datasets were pre-processed using POS Tagging, Stop Word removal, Text Transformation and Clustering (Tyagi & Sharma, 2017). The result obtained from different datasets showed SVM has a higher accuracy on the datasets. However, it failed to use the deep learning methods that would maintain the same level of consistency for the more clustered datasets. Pre-processing of the datasets is always a challenge as it directly reflects the accuracy of the model. If the datasets are large, it can lead to lower accuracy of the model.

2.3 Keyword Extraction

For the extraction of the keywords from a Single Document using Word Co-occurrence Statistical Information (Matsuo & Ishizuka, 2004), the article that was written by Yutaka Matsuo of National Institute of Advanced Industrial Science and Technology and Mitsuru Ishizuka of University of Tokyo. From their research, frequent terms are extracted and a set of co-occurrences between each term and the frequent term. Finding the keyword is the main task of this distribution as the terms are to be distinguished in the document. By calculating the probability distribution of co-occurrence between term and the frequent terms, terms can be distinguished as keywords only when the terms are present on the set of repeated terms. For the degree of biases of distribution, chi-squared is measured. By this same level of evaluation can be done through tf idf as they show certain similarity.To increase the efficiency, several pre-processing techniques like stop words, stemming, PoS tagging are used.

Several processing techniques like new word features are derived according to the knowledge of a document (Xu et al.,) as supplied by Wikipedia. Different processing techniques for keyword extraction and provides significant help for comparison between algorithms. In a document by TF-IDF measures the relevancy of the word in a particular document resulting to query or search different words in a corpus of documents that can be agreeable to use. TF-IDF calculates the inverse proportion of the frequencies of the word for a given document. There is a strong relationship with the document when the TF-IDF values are high which suggests that the words that appeared in the document might be useful and can be used for classification. Before classification, vectorization must be done as the data are present in all natural language sentiments

CHAPTER THREE: RESEARCH METHODOLOGIES

3.1 Research Design

Research design is a conceptual framework within which the entire research will be carried out. It specifies what kind of inquiry the researcher is going to conduct. Selection of an appropriate research design will help the researcher in providing a specific direction for the procedures to be followed to complete the research effectively. Non-experimental / descriptive (cross-sectional) as well as exploratory research will be carried out in this research study (Sileyew, 2019).

Our research design starts with the identification of the problem. Once the problem is identified, different research similar to this problem is searched. A thorough literature review gives the idea about the methodology or steps by which a conclusion of the research can be carried out. To conduct research data is always needed. So, data is gathered from the questionnaires and then subjected towards processing. Then we check whether the data are sufficient or not. If the data is sufficient, we move on to processing while if the data is not sufficient, more literature reviews are done in order to find the alternative way to gather more data. The data source can be both primary and secondary where the primary source of data provides the . After the data is pre-processed, a normality test is also carried out to determine the distribution of our data. After that classification algorithms are done and the accuracy from those algorithms are compared and evaluated. If the accuracy falls below 65%, then pre-processing should again be carried out to refine the data properly.

In our case, we will be also using a confusion matrix for the proper evaluation of those algorithms. Once the evaluation is complete, we can clearly observe which algorithm delivers the highest amount of accuracy, precision, and f1-scores. Based on these analyses, we can conclude that the sentiment analysis is done properly and the results are classified in positive and negative classes with sufficient accuracy.

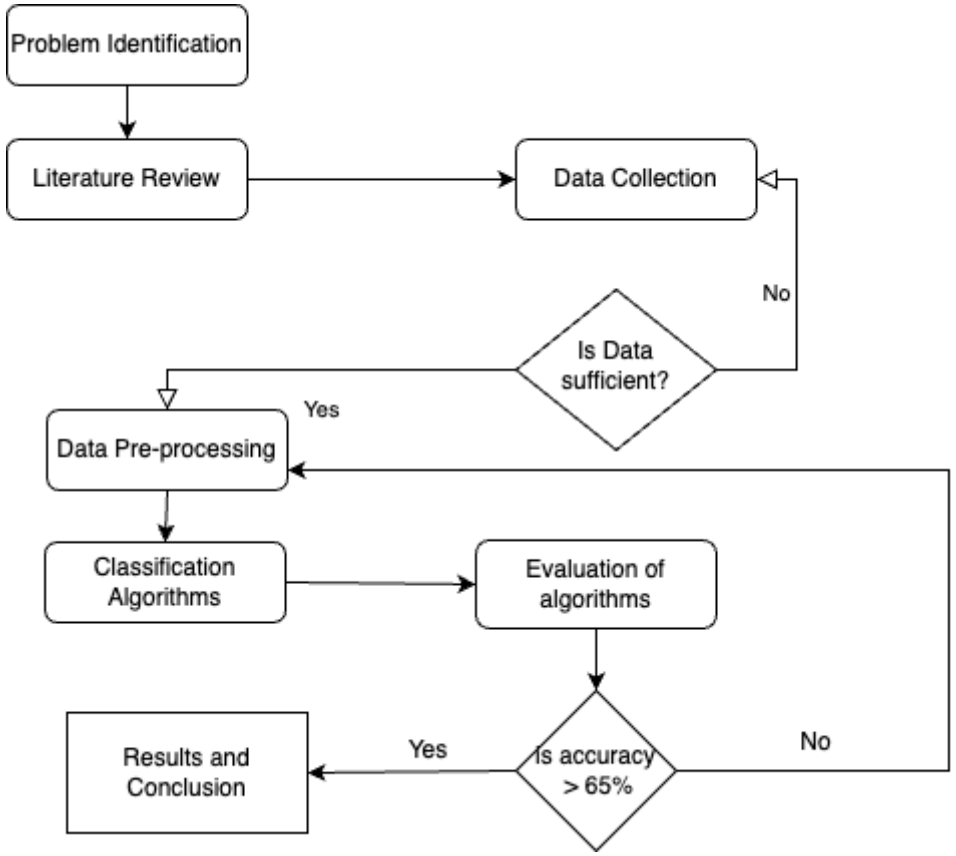

Figure 3.1: Diagram of Research Design

3.2 Sentiment Analysis

To determine sentiments of a text or words i.e. positive or negative sentiments, sentiment analysis is done. The words or texts in the form of natural language are used to analytically determine the appropriate sentiments using natural language processing (NLP) and machine learning (guide).

The main key aspect of sentiment analysis is polarity classification. Polarity refers to the overall sentiment delivered by a certain text or word. Sentiment score is the numerical rating of the text or words often used as polarity. For example, the value of polarity might be greater than 0, less than 0 and equal to 0. If the sentiment score is equal to 0, the text or words of that score can be expressed as the neutral sentiments. Similarly if the sentiment score is greater than 0, the text or words can be expressed as positive sentiments, else all the sentiment scores below 0 are labeled as negative sentiments. This sentiment score or polarity can be calculated for an entire text or just for some few words.

Sentiment analysis uses machine learning and natural language processing (NLP) to identify whether a text or some words contains either negative, positive, or neutral sentiments.

3.2.1 Rule based sentiment analysis

As the name implies, a certain set of rules are introduced inorder to determine the sentiment of the text. This approach includes various NLP techniques like lexicons (lists of words), tokenization, lemmatization and parsing (Bonthu, 2021).

1. The lexicon are lists of positive and negative words introduced by which the impact of different words can be used to describe the sentiment. Positive lexicons might include “good”, “affordable”, and “user-friendly“. Negative lexicons could include “bad”, “pricey”, and “complicated”.

2. The breaking up of texts into small chunks of words like tokens is the process of tokenization. Sentence tokenization splits up text from the sentences like “Today is a good day”, this can be split up as “today”, “is”, “a”, “good” and “day”. Similarly, word tokenization separates words in a sentence.

3. While doing the tokenization, we can remove some things from our text that are not useful for sentiment analysis. Stop words removal is the main process that gets rid of the unnecessary words in a sentence that holds no meaning towards the sentiment of the text like “is”, “am”, “for”, “to” etc. The process to transform the words to their root form can be described as lemmatization. A lemma is the root form of a word like “is, are, am, were, and been” is “be”.

4. The final step parsing is done to count the total no. of positive and negative sentiments present in a text. Some extra rules can be introduced to exactly classify the negative and positive sentiment like “not good” should be classified as negative sentiments. And at last, the total positive and negative sentiments are calculated on the basis of sentiment scores.

3.2.2 VADER Sentiment Analysis

VADER is a lexicon and rule-based sentiment analysis tool that is used to classify the sentiments of a particular text by calculating the polarity of the sentences using the above rule based sentiment analysis. VADER uses a combination of all the sentiment scores present in the text and determines the compound or combined sentiment using the median of those sentiment scores. VADER not only tells about the Positivity and Negativity score but also tells us about by how much the sentiment is positive or negative.

Some neutral structures in sentiment such as “but”, “not” etc can change the polarity of the sentiment. For such measures, VADER uses a special rule that handles the punctuation, capitalization, adverbs and contrastive conjunctions. The below example actually shows how the degree of the sentiments of negative, neutral and positive delivers the compound score of a sentence (Ma, 2020).

Table 3.1: Calculating the sentiment using Vader Classification

Input	negative	neutral	positive	compound
“This computer is a good deal.”	0	0.58	0.42	0.44
“This computer is a very good deal.”	0	0.61	0.39	0.49
“This computer is a very good deal!!”	0	0.57	0.43	0.58
“This computer is a very good deal!! :-)”	0	0.44	0.56	0.74
This computer is a VERY good deal!! :-)”	0	0.393	0.61	0.82

3.3 Machine Learning Sentiment Analysis

Another way to find the sentiment of the text or words is to follow automated sentiment analysis that relies on machine learning (ML) techniques. In this case a Machine Learning (ML) algorithm is trained such that the trained model classifies the sentiment based on both the words and their order. The favorable outcome of the sentiments actually depends on the quality of the training data set which can be obtained from multiple sources and the algorithm that is used to classify the sentiments.

Some other measures to determine the sentiment is to combine both ML and rule-based approaches. By using such approaches, one can get higher accuracy, higher precision but are complex to build as they need certain rules and algorithms both to carry out the classification of the sentiments. Following steps can be used in the ML approach (Medhat et al., 2014).

1. Feature Extraction and Selection
2. Train the data models and Predict the correct classification
3. Use of different classification algorithms like Naive Bayes, Support Vector Machines, LSTM etc. along with the evaluation using confusion matrix or statistical analysis

Above works are done to find the appropriate sentiments of a particular text or reviews that can provide a deeper meaning for a comparative study. Specifically this thesis paper wants to provide the following:

1. Provide basic NLP theories to determine what a text is and what sentiments hold the positive and negative meaning in a word.
2. Explain the two main approaches for the sentiment analysis i.e. Traditional approach or Lexicon Approach and ML approach.
3. Analyze the different output or sentiments obtained by different algorithms.

To conduct a research, certain specific procedures are followed inorder to identify the research topic, selection of the research questions, processing the data gathered, and analyzing the results information about a research topic. In a research paper, the methodology section allows the reader to evaluate a research study and the research design is created to navigate the workflow of the thesis like what steps, algorithms, evaluation techniques are done to complete a research study..

3.4 Sampling

Sampling is the method that allows the researcher to sort out a definite number of samples from among the large population in his field of inquiry. The outcome of any research study is highly influenced by factors like selection of an appropriate sampling method and whether the samples taken from the population is a true representative of the entire

population or not. In the ML approach, the algorithm can work better if the sample datasets are larger. For any kind of ML analysis, a model is created using the available datasets and the model can be much more precise if the number of sample data is higher. For our research purposes, we need a sample of more than 5000 datasets.

3.5 Data Collection

Data collection method is a method employed by the researcher for collecting the data appropriate for the research study. There are two major sources of data collection which are primary dataset and secondary. Primary data collection will be a major source of data collection of this research study where a number of classification models will be created. Secondary dataset are optional in our case as the availability of the secondary data set requires a number of procedures which can or cannot be fulfilled. But mostly secondary data sets are mainly used for the evaluation of the model created using primary data. And the primary dataset for this research is obtained through a questionnaire form where data of around 6000 users are collected.

The fields of the data collected through questionnaire are: Name, Age, Gender, Categories of the product/services, Shopping at, Rating of the product/services, Reviews. The nature of these fields are explained with the help of the below table. Also, the secondary dataset is also taken from a popular E-commerce site named Daraz for the validation of the model. The classification models are created from the survey data which we will refer to as the primary data and the secondary data from the organization will help in the evaluation of the model.

Table 3.2: Collection of the datasets

S.no.	**Name**	**Age**	**Gender**	**Categories**	**Rating**	**Reviews**	**Shopping at**
1	*******	19	Female	Beauty	2	Average priced product	Daraz
2	*******	23	Male	Education	1	Bad services with high price	Daraz

3	*******	26	Male	Electronics	3	Good pricing with different functionalities	Sasto Deals
4	*******	29	Female	Fashion	4	Best product	Gyapu

A google form is created to collect the data of above fields from different no. of users. While the online media is not sufficient, we can collect the data offline through the survey of different people.

3.6 PreProcessing

3.6.1 Sentence Level Processing

In lexical analysis, we already knew the process of tokenization. The tokens can be obtained by splitting the text into words, symbols, and any other meaningful elements that a particular sentence has. Before tokenization, the sentences should be cleaned of items or words that produce no significant role in sentiment analysis. For implementing these steps, regex matching and replacement must be done. These kind of processing may involve the following steps (Singhal, 2020):

- Conversion of the characters into lowercase for consistency:
  While training our model, all consistency should be maintained in the datasets as in the characters of the sentences must be in lowercase. If there are any upper cases available in the sentence. These need to be changed to lower cases.
- Removal of the hyperlinks:
  The hypertext links present in a sentence actually never contribute to the sentiment score of the sentences and hence are removed when encountered.
- Remove tag "@" mentions:
  This symbol "@" is used to tag other people in status or in the document, but this symbol also never contributes to the sentiment score of the sentence and is removed.

3.6.2 Spelling Correction

The given dataset is run through spelling analysis to detect any words that are not present in the dictionary. Since the dictionary can be very large, it is very essential that the checking process be very fast and as much accurate as possible (since we have to train a lot

of datasets of huge size, the training process must be fast enough, also the scoring phase for test data set must be time efficient). False classification of words must be avoided as much as possible.

A spell checker algorithm is used to find the words that are not spelled correctly. Spell checker algorithm not only detects the wrong spelled words but also provides a correct word for the wrongly spelled words.

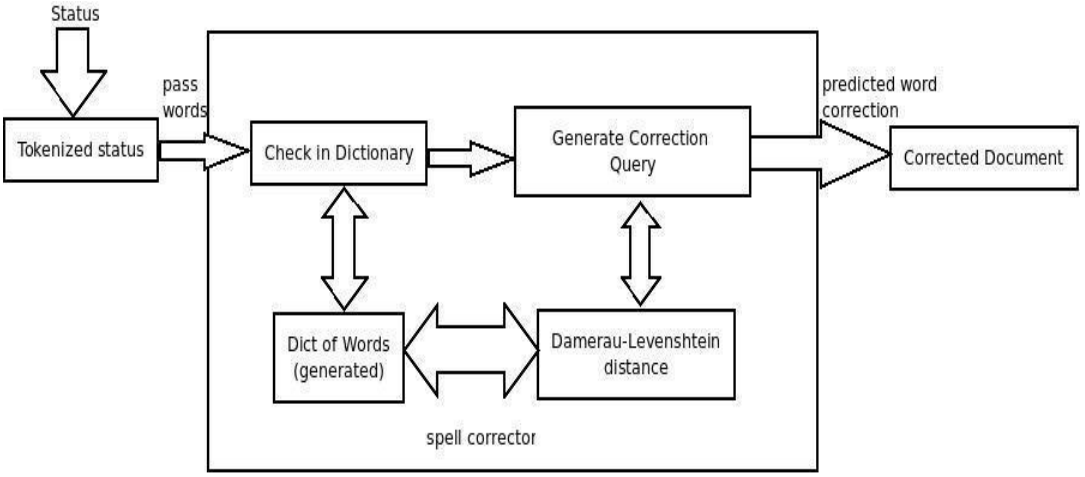

Figure 3.2: Spell Correction Block Diagram

Spell check is accomplished by using the dictionary in python where different packages are available and by which a hash table structure with proper key/values pair is introduced. Dictionary generated is used to check if the words are present or not, providing the word is correct or not. To accomplish the spell corrector, correct words have to be found for the incorrect words. So basically,

Spell Checking Algorithm:

1. Every word in a sentence is checked against a dictionary.
2. If the word is not found in the dictionary, a similar word is suggested from the dictionary to correct the misspelled words.

3.6.3 Token Level Processing

Following the above preprocessing, tokenization is done and the token in status can mean either a normal word, hashtags (#), emoticons or some slang. So different processing of the tokens needs to be done in the following sequence (Menzli, 2022):

- Hashtags (#) and Emoticons:

  If there are Hashtags or emotions (both positive) in a particular sentence, then the sentence yields positive sentiment. So we can replace them with the exact same word without the hash. E.g. #sad was replaced with 'sad'. Same case is valid for negative sentiments.

- Elongated words:

  Sentences can have incorrectly spelled words like "booooks" is mentioned in a sentence for "books".

- Punctuations and additional white spaces:

  Punctuation marks were removed since they don't carry any meaning . E.g: ' Today is a beautiful day! ' should be replaced with 'Today is a beautiful day'. Also some extra white spaces are removed.

- Stop words:

  There are different stop words like the, is, am, are, who, what, when etc that do not contribute to sentiment scores are removed. Different lists of stop words were obtained from the popular NLTK package through python and were used as a dictionary to remove such stop words that do not hold any meaning to the sentences.

3.7 FEATURE EXTRACTION

3.7.1 TF-IDF

Up till now, the preprocessing like stop-words, spelling correction etc are done. Now, the term frequency and the inverse document frequency of the document are measured. TF (Term Frequency) measures the number of times a term (word) occurs in a document. Due to the presence of documents of much larger size, normalization must be done before finding the term frequency which is achieved by dividing the term frequency by the total number of terms (*Tf–idf*).

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{max\{f(w,d):w \epsilon d\}} \quad (3,1)$$

The frequency of the term helps in finding the relevancy of the document. IDF also known as Inverse Document Frequency is calculated to determine the weightage of the less frequent terms

$$idf(t,D) = log\ \frac{N}{|\{d \epsilon D: t \epsilon d\}|} \quad (3,2)$$

Where,

N represents total number of document within the corpus

$|\{d \in D : t \in d\}|$ = the number of documents where the term appears (i.e. t f (t, d)! = 0).

Up till now, the individual calculation of TF-IDF is done and combining those two we can get the value of TF-IDF which can be expressed in equation as:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \quad (3,3)$$

Higher the TF-IDF value, higher the relevance of the document. The main idea behind this is the log function that is used to calculate the idf value which in response affects the tf-idf. If the ratio inside the log becomes closer to 1, then TF-IDF value moves closer to 0.

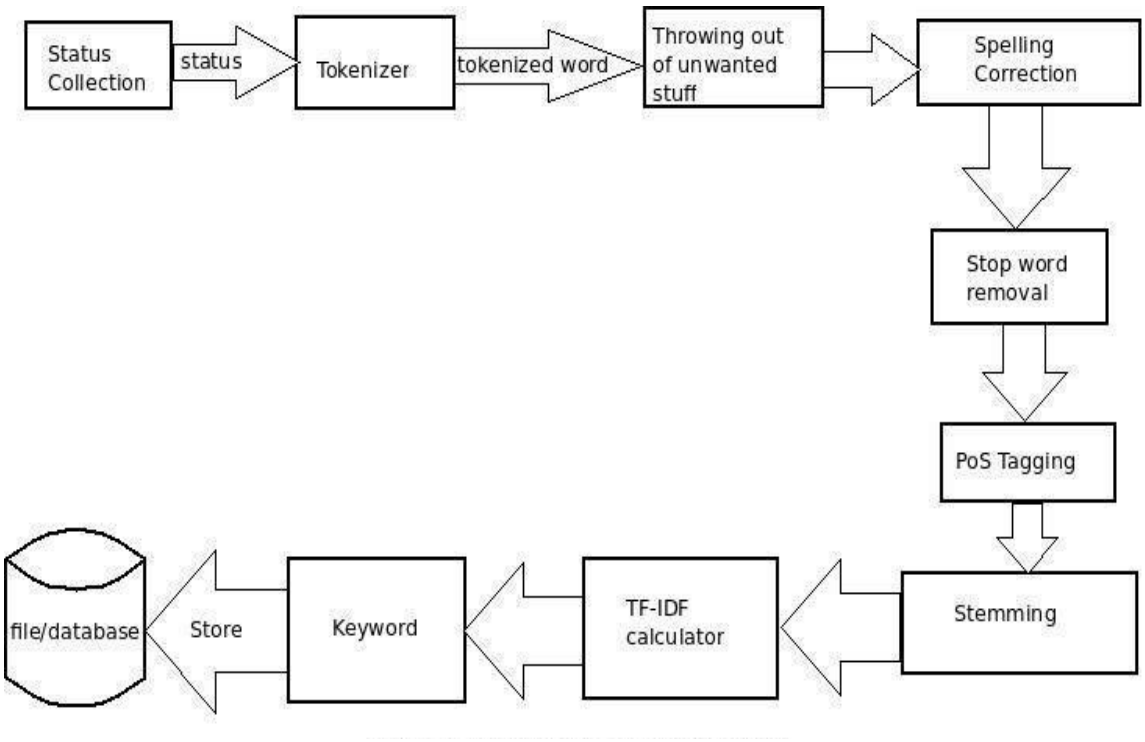

Figure 3.3: Preprocessing for keyword extraction

3.8 Model Creation

The primary goal of this thesis is to perform sentiment analysis of ecommerce data using machine learning algorithms. First the data is gathered from the questionnaire targeting different categories of the products from different e-commerce platforms. Then the data are

subjected through some preprocessing before the use of machine learning algorithms (Garcia et al., 2020).

The preprocessing includes text separation using tokenization and sentiment detection using Bag-of-words or TF-IDF method which will be sent for classification. The sentiment classification includes different methods like SVM, Deep Learning and Naïve Bayes to determine the exact class of the sentiment words i.e. Positive, Negative, or Neutral.

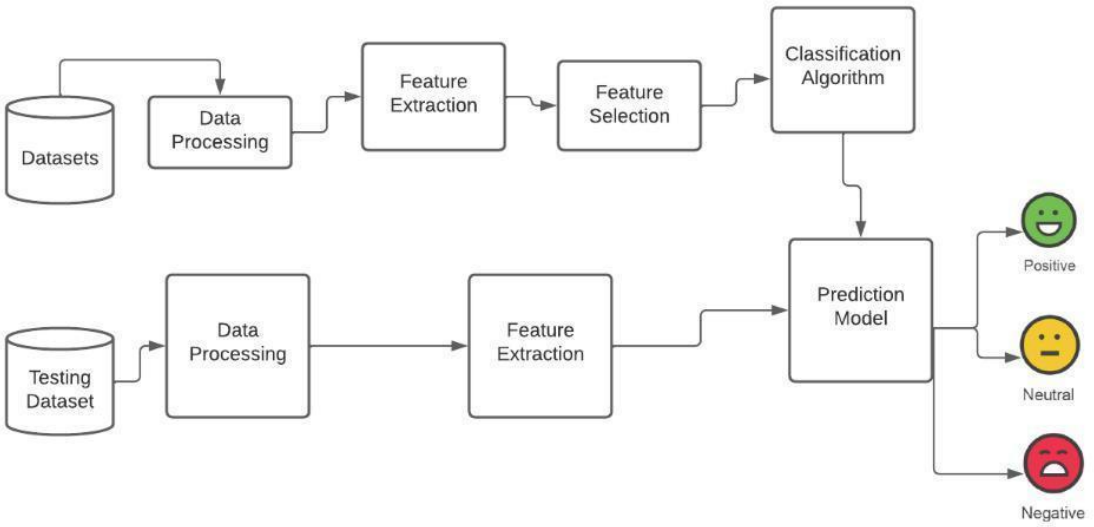

Figure 3.4: Block Diagram of system classifier

3.8.1 Naive Bayes

Naive Bayes algorithm is a probabilistic learning algorithm mostly used in Natural Language Processing (NLP) to find the sentiments of the text, words etc based on the probabilistic events. The algorithm is based on the Bayes theorem which predicts the probability when there is a occurence or probability of certain events (Ray, 2017).

Bayes Theorem can only determine the probability when prior knowledge is available. It is based on the following formula:

$$P(A|B) = P(A) * P(B|A)/P(B) \quad (3,4)$$

where,

P(B) = prior probability of event of B

P(A) = prior probability of class of A

P(B|A) = conditional probability of B given class A probability

3.8.2 LSTM (Long-Short-Term_Memory)

LSTM networks are a special type of RNN that uses more special units to hold the data for a longer period of time. As the name implies, LSTM units include a 'memory cell' that can maintain information in those memory cells for a long period of time for longer term dependencies This memory cell lets them learn longer-term dependencies. Like any other RNNs, LSTM also includes a series of hidden layers that maintains the data optimized and sent to the correspondent hidden layers, continuously to the output layers where the optimized output can be discovered (Graves, 2015).

Almost all recurrent neural networks have the form of a chain of repeating modules of neural network where the outputs are again provided to the input layer for the improvisation of the output.

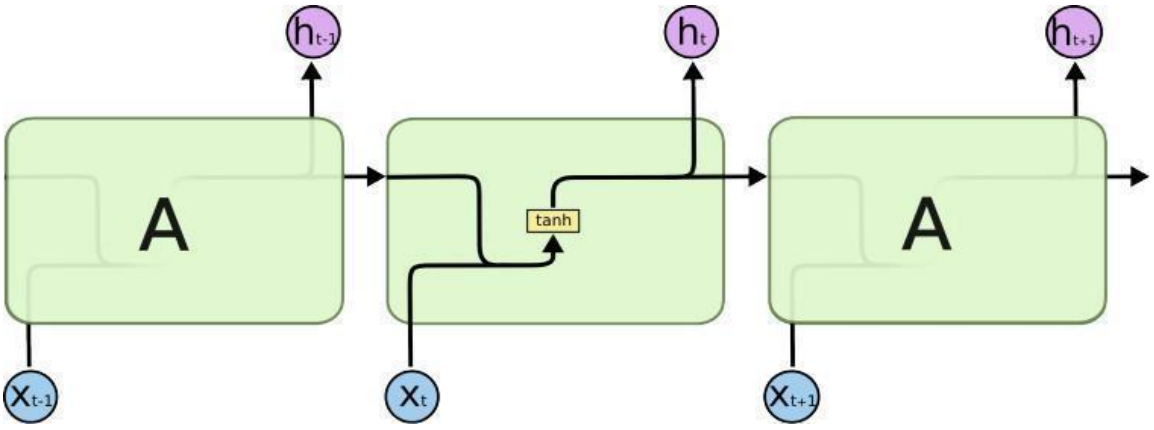

Figure 3.5: Repeating RNN model (Vijayprabhakaran & Sathiyamurthi, 2020)

The LSTM model is a special type of RNN that can hold values for the future and transfer it again to the activation function for better optimized results. The maximum words allowed in a sentence is 128 while there is only one dense layer. However, there are 100 LSTM hidden layers which provide a good modeling when the train and test data are separately sent to the LSTM model. LSTMs also have this chain-like structure with multiple hidden layers all connected with the input of the other layer that stores the previous value in a cell comprising the output to be optimized, but the repeating module has a different structure. Every hidden layer in LSTM acts as a memory buffer where the previous values are saved for future use as the constant back propagation input is required for the optimization. A recursive neural network is created by applying the same set of

weights recursively over a differentiable graph-like structure by traversing the structure in topological order. Such networks are typically also trained by the reverse mode of automatic differentiation. They can process distributed representations of structure, such as logical terms. A special case of recursive neural networks is the RNN whose structure corresponds to a linear chain. Recursive neural networks have been applied to natural language processing. The Recursive Neural Tensor Network uses a tensor-based composition function for all nodes in the tree.

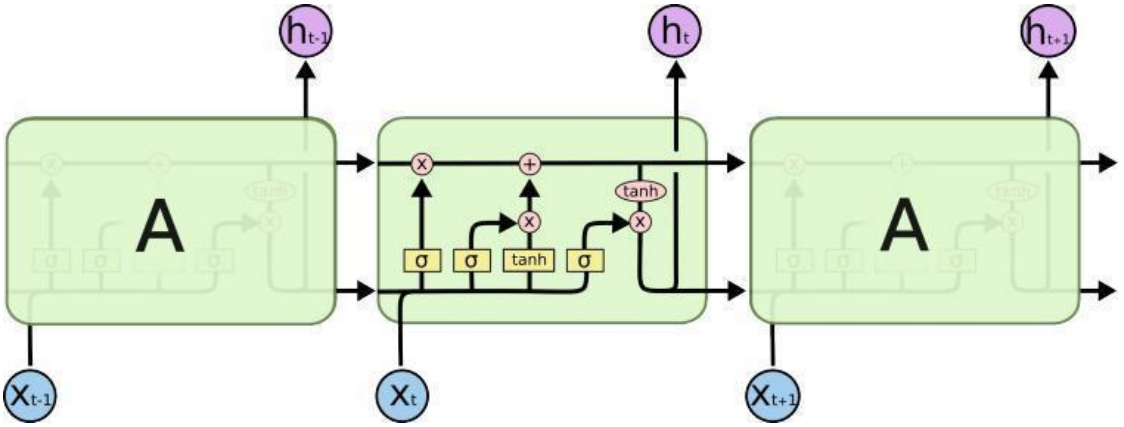

Figure 3.6: Repeating LSTM module with 4 interacting layers

Let X_t be the input and h_t be the output at time t. At time t, the equation of gates, memory cell, input and output of the LSTM cell are as follows:

$$i_t = \sigma(W_i[x_t] + R_i[h_{t-1}]+b_i \quad (3,5)$$

$$f_t = \sigma(W_f[x_t] + R_f[h_{t-1}]+b_f \quad (3,6)$$

$$o_t = \sigma(W_0[x_t] + R_0[h_{t-1}]+b_0 \quad (3,7)$$

$$g_t = \tanh(W_x[x_t] + R_x[h_{t-1}]+b_x) \quad (3,8)$$

$$c_t = f_t * c_{t-1} + i_t * g_t \quad (3,9)$$

$$h_t = o_t * \tanh(c_t) \quad (3,10)$$

where i_t, f_t, o_t indicates input gate, forget gate and output gate respectively and c_t defines the memory cell to store the past state. The g_t and h_t refers to input and output of the LSTM cell respectively and h_{t-1} refers to the output of the previous LSTM cell. Sigmoid and tanh are the activation functions to map the non-linearity (Vijayprabhakaran and Sathiyamurthi, 2020).

In the standard LSTM network, sigmoid is used as the gating function and the tanh is used as the output activation function. The difference between a RNN model and LSTM model is that RNN provides the activation function and back propagation only whereas the LSTM model provides the same functionality as the RNN model as well as the provision to store the data for a while to compare and optimize the activation function inside the model.

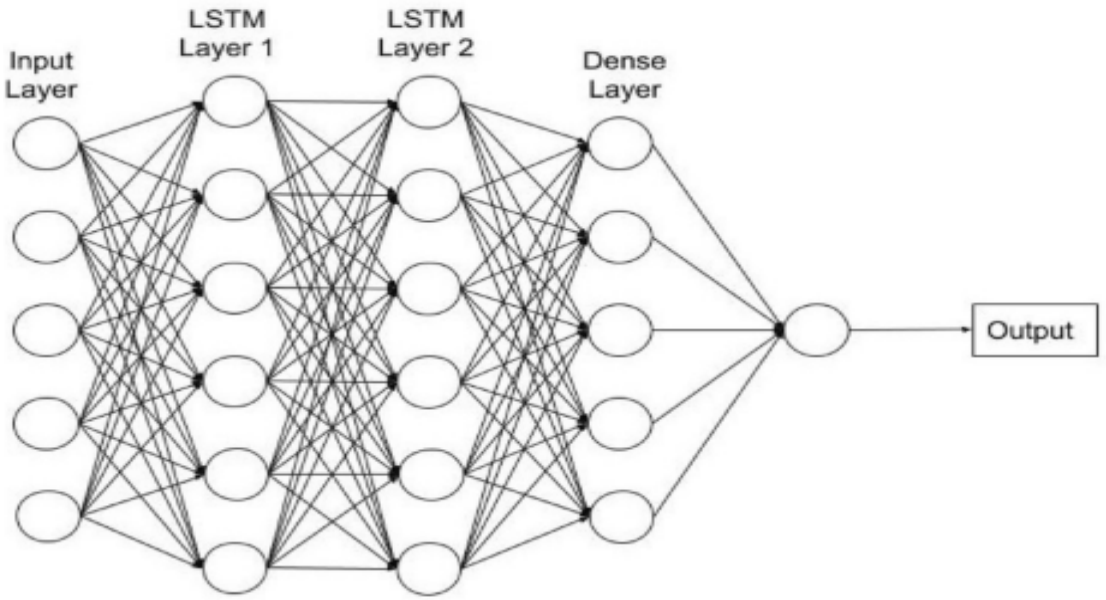

Figure 3.7: LSTM network model

The LSTM model mainly consists of input layer, LSTM layer1 and layer 2 which are also known as hidden layers and Dense layers. For the part that we use to run the data set into the given LSTM model, we provided our dataset as the input to the input layer and two 176 hidden layers were set up in the model so that our model works with much more precision and optimizes the output. After that 1 dense layer was introduced which subjects the output layer.

3.8.3 Support Vector Machine

Support Vector Machines (SVM) are generally used for both classification and regression tasks. Since the method is supervised the outputs are already known and using the

classification techniques such data are used to train the SVM model which then classifies the unknown data with maximum accuracy and precision.

Given a set of training data, the goal is to find a function f (x) that has at most ε deviation from the actual targets of the training dataSince the deviation is only allowed to ε, some errors are permitted to the deviation of ε (Smola & Scholkopf, 2002). Following is the linear function that can be described as:

$$f(x) = \langle w, x \rangle + b \tag{3,11}$$

where w ∈ X, b ∈ R, . denotes dot product in X. the regression function of the SVM will use a penalty only if the predicted value f (x) is more than ε distance away from the actual value y_i.

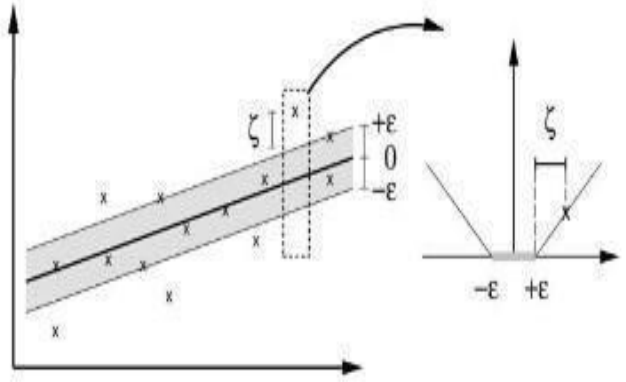

Figure 3.7: Soft Margin loss setting in linear SVM (Smola & Scholkopf, 2002)

The given equation can be solved by dualization using Lagrange multipliers. The final function f(x) is expressed below

$$f(x) = \sum_{i=1}^{k} (\alpha_i - \alpha^*_i)\,(x_i, x) + b \tag{3,12}$$

where α_i and α_i^* are Lagrange multipliers. We can see that w in the equation 7 can be specified by the linear combination of training examples x_i.

It follows that for all samples inside the ε tube Lagrange multipliers α_i and α_i^* are zero whereas samples outside the ε tube have non zero coefficients. Hence in order to define w we only need non vanishing coefficients which come from samples outside the ε tube, which are the Support Vectors. With the help of these support vectors we divide a hyperplane that separates the two different classified datas. The main idea behind the SVM is to create a hyperplane separating the different features of data and clustering the same features of data in a particular plane. The classification involves multiple hyperplanes that separate the datasets.

3.9 Normality Test

Normality test is usually carried out to determine whether the data set follows normal distribution or not. There are many tests involved in finding the normality. Before performing a normality test, some graphical methods are also used to visualize whether the distribution is normal or not. Usually three types of graphical tests are used to visualize whether the datasets are normally distributed or not. These graphical tests include Q-Q plot, Box-plot and the histogram plot. From these observations, one can visualize whether the distribution is normal or not. For the analytical test, we have Kolmogorov-Smirnov (K-S) test, Anderson-Darling test etc. The Anderson-Darling test makes use of the specific distribution in calculating critical values and a series of other values where we can see test of significance for different alpha values. The main idea behind the Anderson-Darling test is that if the p-value is less than 0.05, we reject the null hypothesis i.e. Our datasets do not follow normal distribution.

Test Statistic: The Anderson-Darling test is defined as

$$A^2 = -N - S \tag{3,13}$$

Where

$$\sum_{i=1}^{N} \frac{2i-1}{N} [lnF(Y_i) + ln(1 - F(Y_{(N-i-1)}] \tag{3,14}$$

Following are the two sets of hypothesis for our normality test

H_0 = Our dataset follows the normal distribution.

H_1 = Our dataset does not follow the normal distribution

If the data does not follow normal distribution, there are different techniques with which we can make our distribution normal. The best method is to use the log transformation when the data are skewed and kurtosis is high, the log transformation of the data helps to arrange themselves into normal distribution.

Anderson Darling Test may or may not always verify the normal distribution, so graphical analysis must be done to showcase the actual distribution of the data. The different methods for graphical analysis of the normal distribution are box plot, Q-Q plot, and histogram plot. So, one can use both analytical and graphical tests to ensure that the distributions are normal.

CHAPTER FOUR: RESULTS AND DISCUSSION

4.1 Study of the data and Sentiment Analysis

4.1.1 Study of the dataset

Before Classification techniques, datas is visualized and the nature of the datas is studied. Following figure 4.1 shows the data analysis of the no. of users shopped at different E-commerce sites. During the survey, we enlisted four major ecommerce platforms from which Daraz has more no. of reviews which indicates that people are mostly engaged in Daraz than any other e-commerce sectors for online shopping.

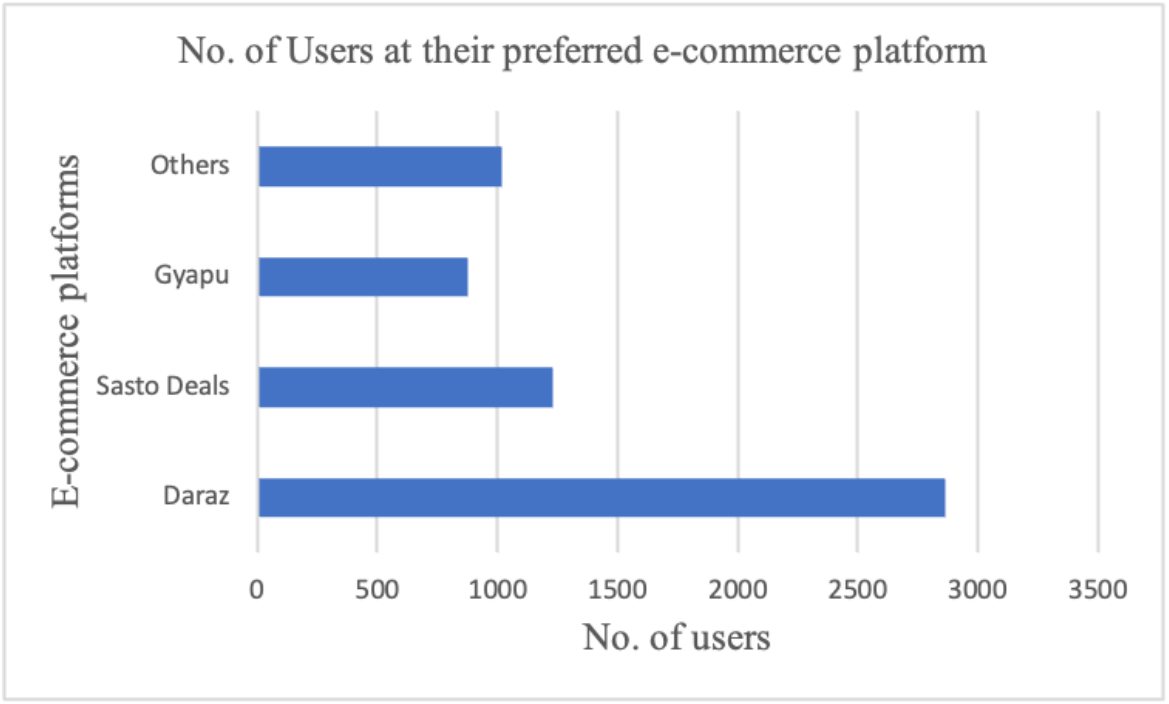

Figure 4.1: No. of users shopped at different ecommerce sites

While the majority of the users are deviated towards Daraz, other ecommerce sites are also seen with some user's activities on their products. However, those other ecommerce sites need to upgrade their services as the customers are flocking towards Daraz. Following figure shows the data analysis of the user's satisfaction towards a product or services they used in Ecommerce. Rating 1 shows dissatisfaction and poor reaction to the product/services while Rating 4 shows satisfaction and good reaction to the product/services.

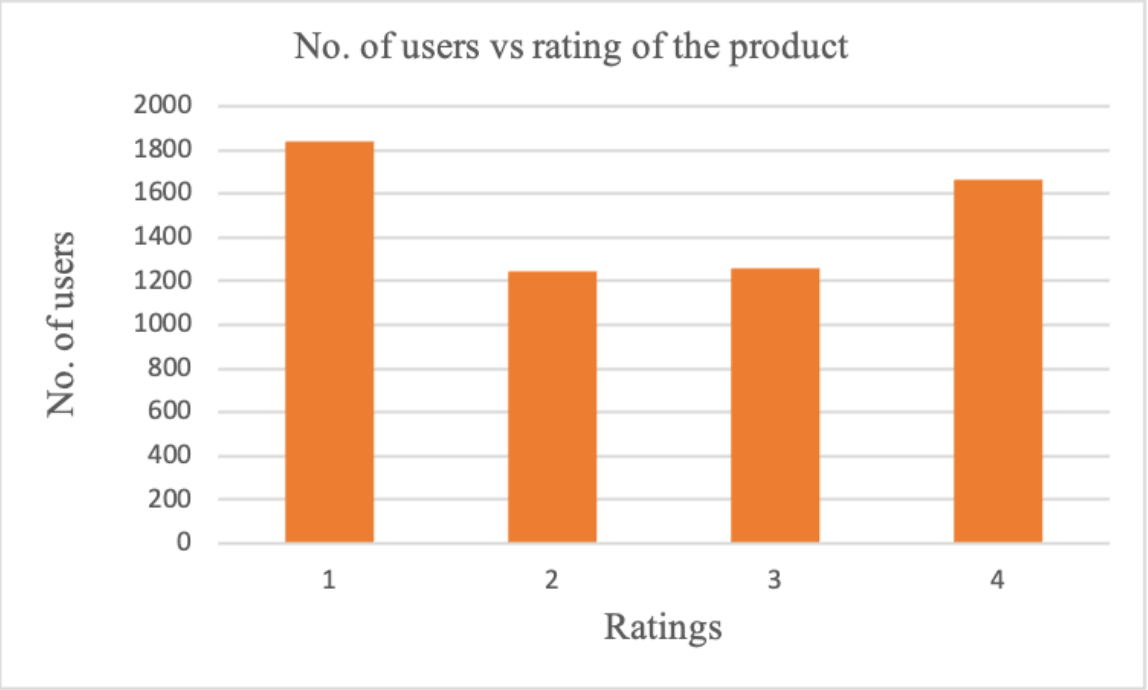

Figure 4.2: Rating of the product or services vs Users

From Figure 4.2, we can see that the rating of the products or services used in Ecommerce are almost equally rated from low to high where 1 being low and 4 being high.

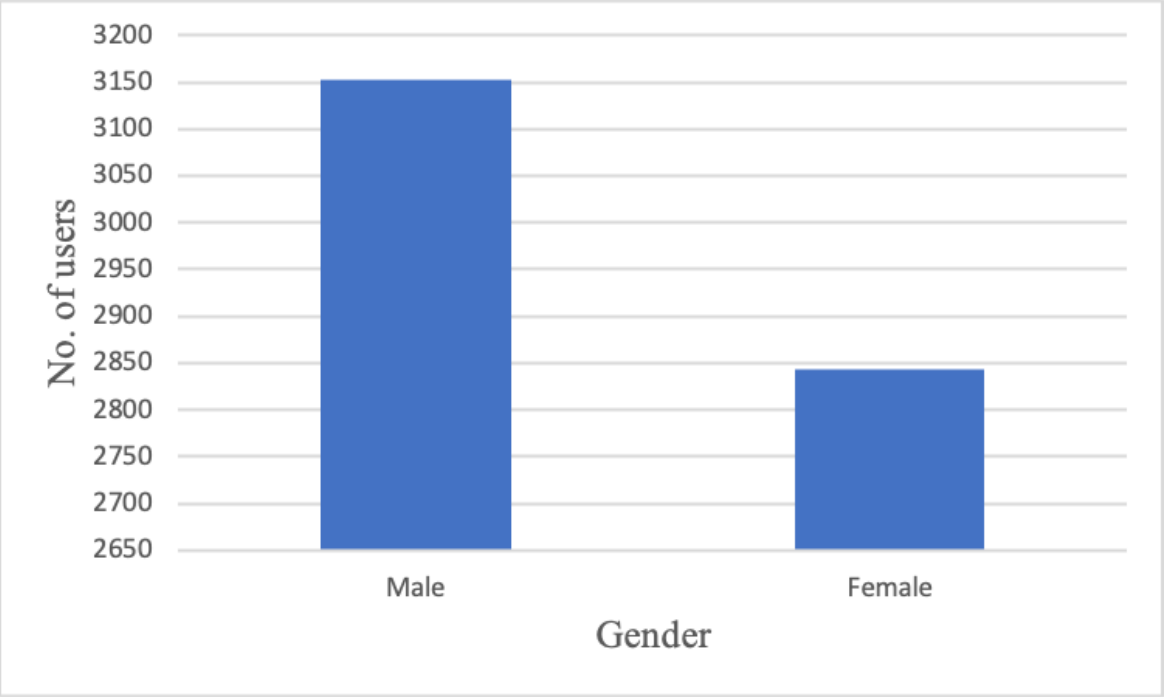

Figure 4.3: Male vs Female users in the dataset

From this we can interpret that there are plenty of users who are satisfied, dissatisfied and ok with the products or services in the Ecommerce platform. Following figure shows the gender specification of the people involved in Ecommerce products/services.

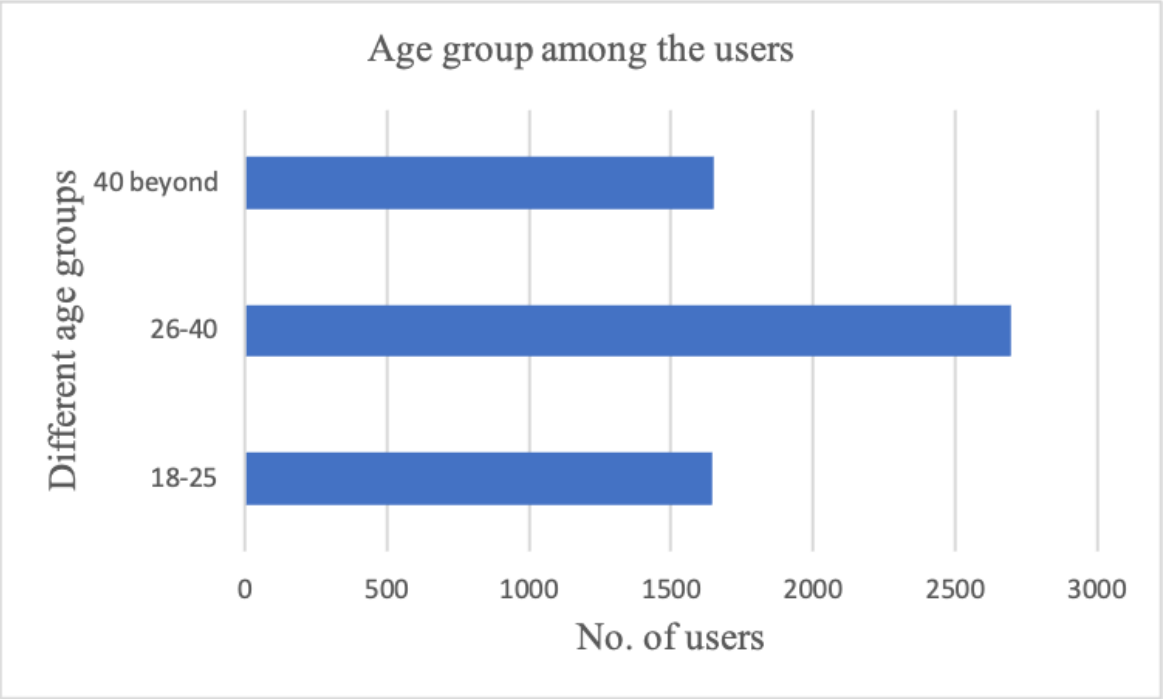

Figure 4.4: Users with different age group in the dataset

From figure 4.4, we can see that there are more male users than female users who are participating in the Ecommerce products/services of Nepal. Similarly,Following figure shows the age distribution of the users who are involved in Ecommerce products/services. We divided the total age groups in three categories: 18-25, 26-40 and 40 beyond. This is done to find the density of certain age-group users participating in Ecommerce products and services. From this we can have an understanding that people with the age group 26-40 are much more engaged in e-commerce. These are the age groups who are mostly involved in the employment sectors and are contributing to the new technology incorporation. As the new innovation is found, these age groups are mostly active in finding the features of a particular innovation. Also, they are the most busiest age group as they are mostly active in part-time and full-time employment which gives them less time to shop which in return increases the involvement of the e-commerce services. Also, most of these age group people tend to buy the product only after the confirmation of the product so e-commerce is the most ideal platform as all the details about the product, their services, their features are

clearly available on the e-commerce platform along with the provision to compare the same products of different brand values. From figure 4.4, we can observe the distribution of users among different categories. We can visualize that the age group of 26-40 are actively participating in the e-commerce sectors. Since Nepal is improving its e-commerce sector, we can see that other age groups are also actively participating in the e-commerce sectors.

After the visualization, datasets are sent for pre-processing where the datas are pre-processed and sent for classification between the two classes i.e. Positive class and Negative Class using different classification algorithms. The main goal of the classification algorithm is to classify whether the data sets belong to the positive class or the negative class.

4.1.2 Word Cloud of the positive and negative sentiments

Word clouds are generated to feature the frequent words that have been used in the reviews. The continuous occurrences of the frequent and impactful words are shown in a figure with different text format and styles such that the particular word cloud can be easily visualized by any user. Word clouds mostly feature the frequent words involved but in case of sentiment analysis, it also provides the analytics that which word made the most impact to label a particular data or reviews in either positive or negative class.

The dataset was pre-processed using various algorithms and the data were subjected to a classifier which provides the result of negative and positive sentiments of the product reviews. Based on the negative and positive reviews, the datasets are divided into separate arrays and the word cloud is created based on the result. Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the document. This type of visualization can assist evaluators with exploratory textual analysis by identifying words that frequently appear in a set of interviews, documents, or other text. Word clouds can also be described as the most effective analysis in the sentiment classification as the words can be directly visualized in the word cloud of both positive class and negative class.

Figure 4.5: Positive Word Cloud obtained from the reviews

From the data in the above figure, we can see the repetitive words in big letters whereas the least frequent words are showcased in small letters for positive class. Using the library of python, we introduced our positive class data into the framework which then exhibits the above word cloud. These words in the above word cloud are the main factors behind establishing any sentence into the positive class. The use of such words in the sentence gives it much more probability to classify in the positive class. The words in the above figure clearly shows that the repeated words signify some positive sentiments in a text review/sentence. After the pre-processing, only valid data remains which contains the negative, positive and neutral meaning. Setting aside the neutral and negative sentiments, the positive sentiment class is used and the words used in them are portrayed in the above word cloud. We can clearly see the words like good, ready, great, satisfaction etc. which are the words that resemble the positivity in a sentence or text. The occurrence of such words in a text or sentence makes the sentence move towards positive polarity. The more the positive polarity, it can be classified in the positive sentiment class.

Figure 4.6: Negative Word Cloud obtained from the reviews

From the data in the above figure, we can see the repetitive words in big letters whereas the least frequent words are showcased in small letters for negative class. Using the library of python, we introduced our negative class data into the framework which then exhibits the above word cloud. These words in the above word cloud are the main factors behind establishing any sentence into the negative class. The use of such words in the sentence gives it much more probability to classify in the negative class. While the use of word clouds of both positive and negative sentiments provided the details of the data collected from the survey for 5996 users, we can clearly see the kind of words used by the users to interact with the e-commerce platform to share their views/ opinions about a particular product they buy online. The user's response about the e-commerce product/services can be good or bad. The good responses are displayed by the positive word cloud where the users implemented many good words that impacted the response to fall into the positive class. Same goes for the bad responses which are displayed by the negative word cloud where the users embark their frustration with the negatively impacted words into the reviews making them fall into the negative class.

4.1.3 Normality Test

Both the analytical and graphical test for normality are conducted where for the analytical test, Anderson-Darling Test was used for normality testing where the null hypothesis is created that the distribution follows normal distribution and the alternate hypothesis is that it does not follow normal distribution. According to the p-value, the null hypothesis is accepted and rejected. If the p-value from the Anderson-Darling test is less than 0.05, then we reject the null hypothesis else we accept the null hypothesis. And for the graphical test, Histogram plot and Q-Q plot is done providing information that the datasets follow normal distribution or not.

From the data gathered, all of the sample data were subjected to mathematical calculation of Anderson-Darling by which p-value is measured. And the datasets were subjected for both histogram plot and Q-Q plot to showcase the normality distribution. For the first graphical test, the histogram plot was done and our distribution looks like the following figure.

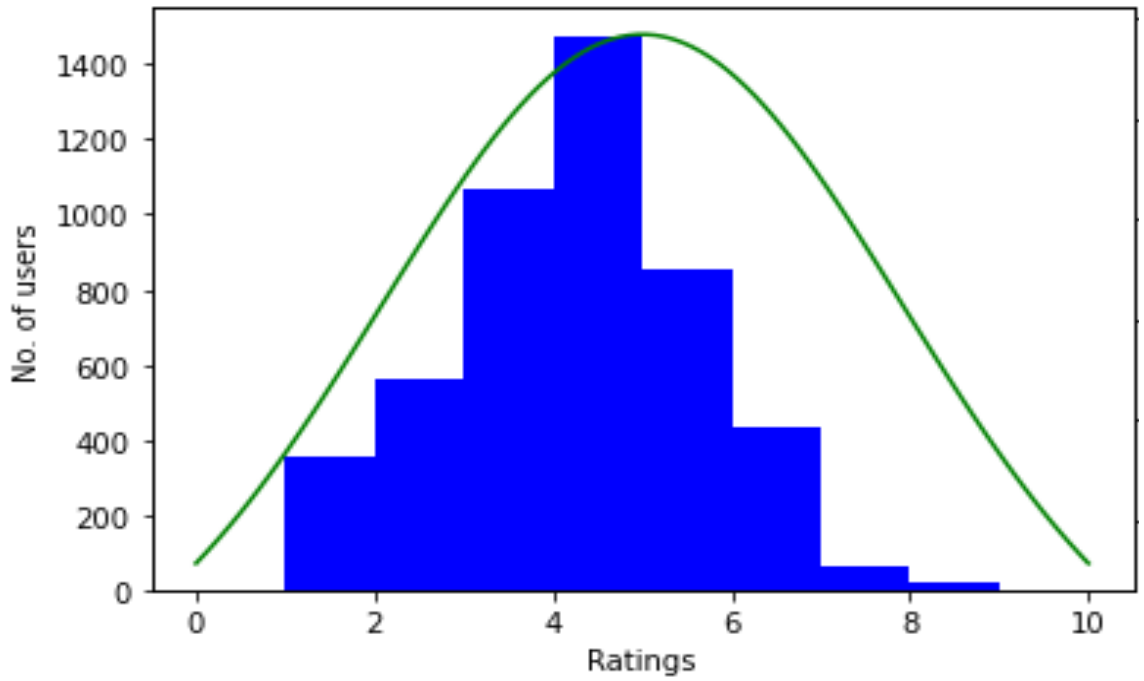

Figure 4.7: Histogram plot showcasing normal distribution

Similar Q-Q plots were created for the second graphical test, which determines if a random variable has a normal, uniform, or exponential distribution by looking at its data.

It is obvious that a distribution is a normal distribution if all of the points shown on the graph exactly lie on a straight line since it is evenly aligned with the standard normal variate, which is the basic idea behind the Q-Q plot. Normally, histogram plot is enough to showcase that the datasets represent normal distribution. But in some cases, we need to be sure so that a graphical and analytical test must be done to validate that the given dataset follows normal distribution. To be sure another graphical plot named Q-Q plot was done just in case to show the normality distribution of the data. With the right data visualization from the histogram and Q-Q plot, we can say that the graphical test shows that the data are normally distributed. However, analytical tests are also done to ensure that datasets are in normal distribution.

The following figure was obtained from the Q-Q plot explaining that data are normally distributed.

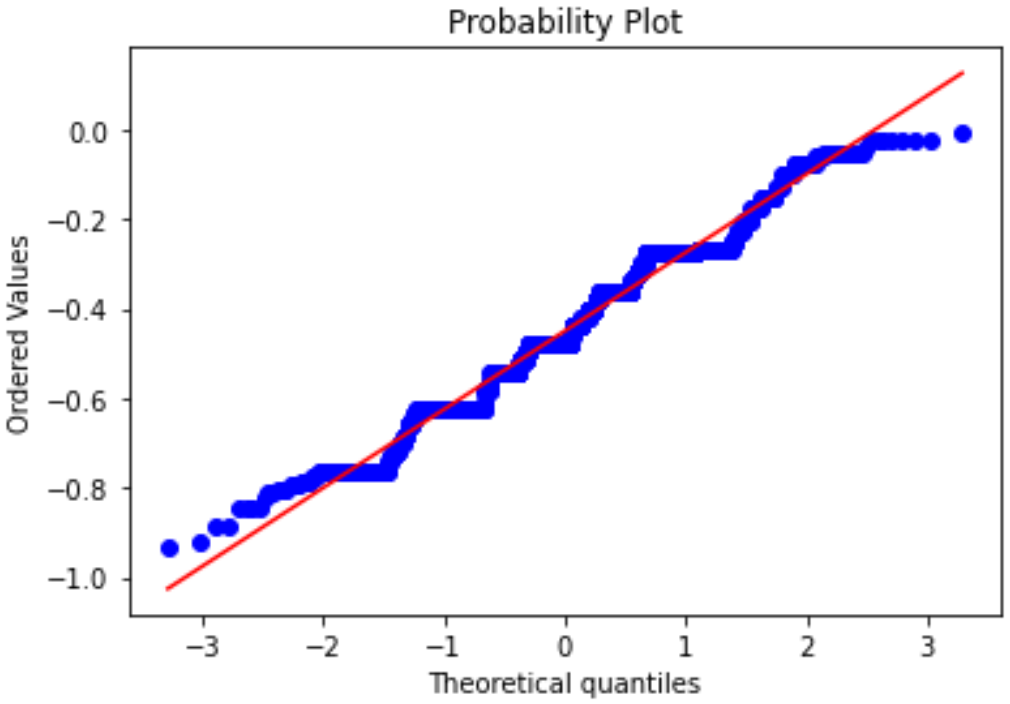

Figure 4.8: Q-Q plot for normality testing

To check normality, Anderson-Darling test was conducted on the independent variables. The A-D hypothesis states that:

$H_0 : \overline{s}$ = Normal distribution

$H_a : \overline{s} \neq$ Normal distribution

The Anderson-Darling test p-value was found to be 0.452 which is higher than 0.05 (when α =0.05) indicating statistical difference between the sample and normal distribution. then we can say that the sample follows a normal distribution.

4.1.4 Sentiment Analysis

A total of 5996 users' responses were recorded from the questionnaire/survey. The data includes the following fields: Name, Age, Gender, Categories, Shopping at, Rating, Reviews. This data is used to model the different classification algorithms. The sentiment analysis result was found to be no. of positive sentiments much more greater than negative sentiments.

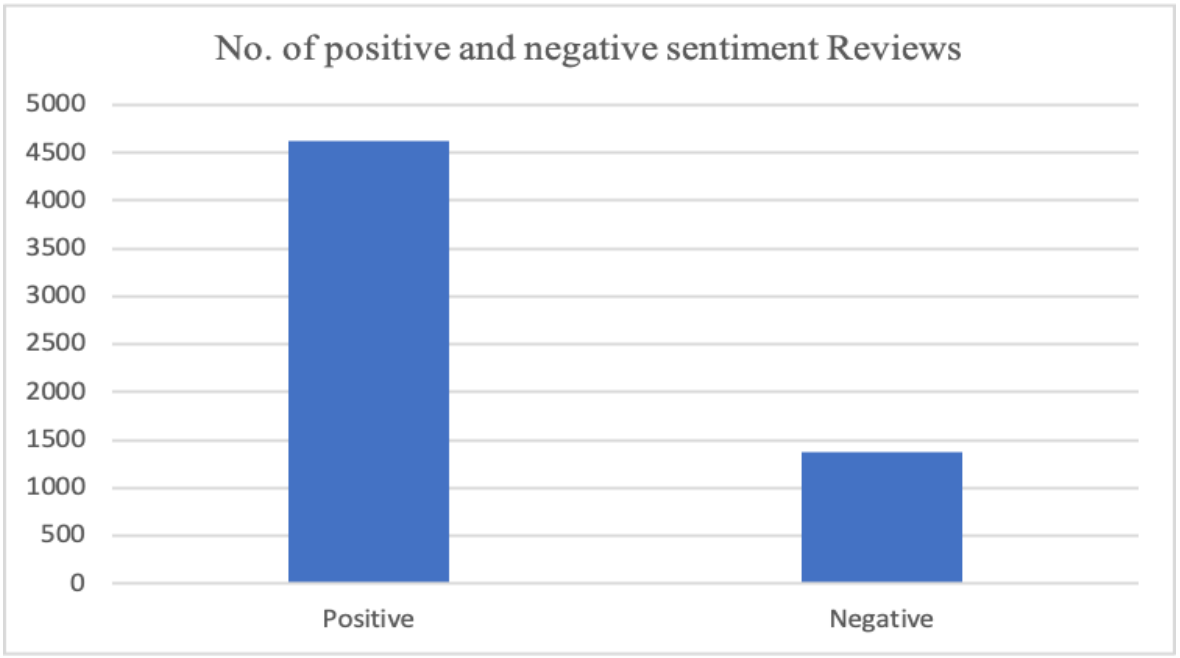

Figure 4.9: Sentiment analysis of the total users

From figure 4.9, we can clearly see the main output is that the classification of the total no. of users are done and users are divided into positive and negative classes based on their sentiment scores. If the sentiment score is greater than 0, it belongs to the positive class and if it is less than 0, it belongs to the negative class. Also if the sentiment score is equal to zero, it belongs to the neutral class. For the better analysis for finding the good and bad sentiments in a review, neutral reviews are discarded in our case to emphasize more on the reviews that make a solid impact on the e-commerce organizations.

4.2 Comparison of sentiment scores of different e-commerce platforms

Another study here provides the data of users in different E-commerce sectors bearing different numbers of reviews. The data is displayed in the table below:

Table 4.1: Sentiment analysis result of different e-commerce sector

	Positive	Negative
Daraz	2233	634
Sasto deals	915	316
Gyapu	671	203
Others	180	40

From the data obtained after using VADER classification, the number of positive sentiment and negative sentiment of the data were visualized for all different sectors of E-commerce in Nepal like Daraz, SastoDeals, Gyapu, Others. Daraz has the lead in the highest number of positive sentiments with their products/services when compared with other e-commerce sectors.

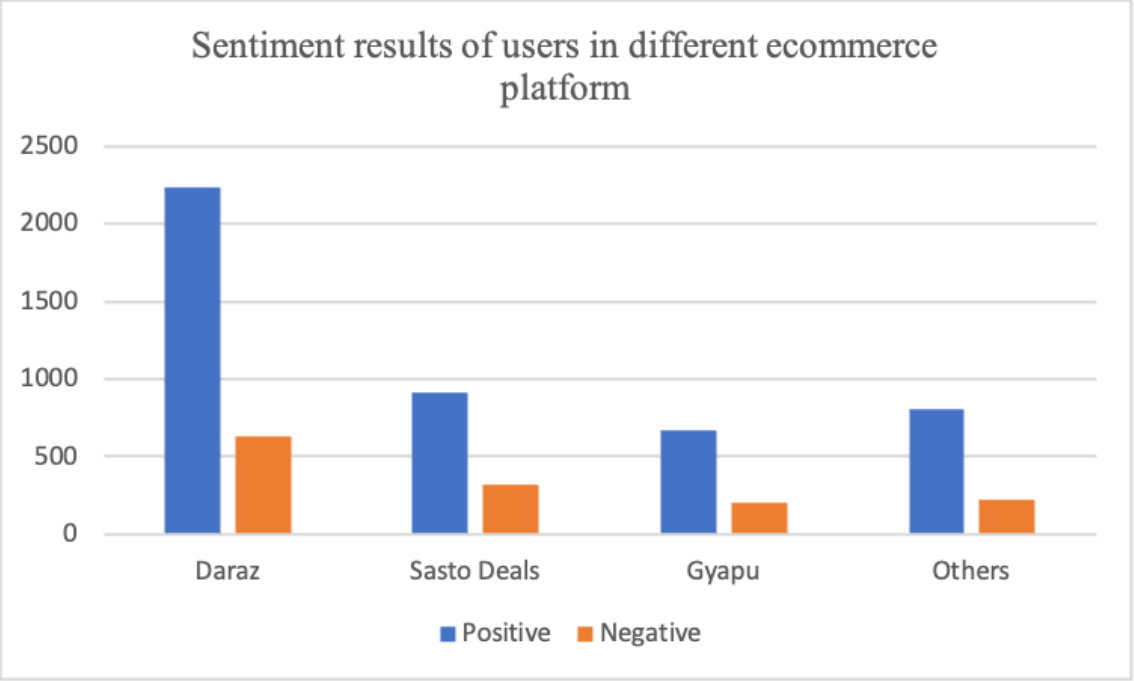

Figure 4.10: Users data with sentiment across different e-commerce

All the data obtained were pre-processed using Stop words, Tokenization, Lemmatization and the word cloud is created for the positive and the negative reviews

data. Positive word cloud is created using the highest rating reviews. The word cloud displays the most frequency word appeared in reviews of the rating 4 product/services.

We have used 4 algorithms for the study of our sentimental data. These are Vader Sentiment Classification, Naive-Bayes Classification, SVM classification and LSTM model for sentiment classification. All the algorithms were implemented and for the evaluation, we have created the confusion matrix and the classification report for all algorithms. From the provided table below, we can clearly analyze that the accuracy among the given models is high for the SVM classifier with an accuracy of 98%. While the least accurate model was for VADER classifier with 68% accuracy. While using different algorithms, we get different types of results. The evaluation metrics of Machine Learning algorithms can be expressed by creating a confusion matrix and calculating the precision, recall and F1-score along with the accuracy of the model.

4.2.1 Analysis of the classification algorithms

For the first classification, we used the VADER classification techniques that exhibited a model which provided the positive and negative class of the datasets. The data were sent inside the model and for the evaluation and a label was introduced on the basis of ratings of the product/services which was used for the evaluation of the result obtained from the classifier. The following results were observed for VADER sentiment classifiers which showcase that the model gives 68% accurate results with 0.20 f1-score for negative class and 0.80 f1-score for positive class. From this result we can say that model performance for positive class was ok whereas model performance for negative class was not good according to the f1-score of the both classes.

For the second classification, we used the Naïve-Bayes classification techniques that exhibited a model which provided the positive and negative class of the datasets. The data were sent inside the model and for the evaluation and data sets were divided into training and testing datasets. The Naïve-Bayes model was trained using training data and test data were used to evaluate the model created from training data. The following results were observed for Naïve-Bayes sentiment classifier in the form of confusion matrix which showcase that the model gives 68% accurate result with 0.12 f1-score for 0 or negative class and 0.80 f1-score for positive class. From this result we can say that

model performance for both positive and negative classes was excellent according to the f1-score of the both classes. The different evaluation of the algorithms along with their precision, recall and f1-score are presented in the following table for both positive and negative classes. These calculations are done by first determining the confusion matrix for each algorithm and calculating these scores using the obtained confusion matrix.

Table 4.2: F1-score evaluation for VADER, SVM, Naive Bayes

Model	**F1-score for negative class**	**F1-score for positive class**
VADER	0.20	0.80
Naive-Bayes	0.12	0.80
SVM	0.15	0.82

Separate tables were created to display the classification report obtained from the confusion matrix for positive class and negative class. Similarly, another table is created to compare the accuracy result from the different classification models used. From Table 4.2 and 4.3, f1-score is widely used to evaluate the model. If the f1-score is high then the model executes best performance for that dataset, similarly lower f1-score exhibits the low performance of that model on the dataset.

Another table 4.4 provides the data of model accuracy. Different classification algorithms are used out of which, the LSTM model exhibits the highest accuracy. So, we can say that for the e-commerce reviews data, the best accurate model is the LSTM model compared to other classification models. While the other models are in close competition when dealing with the deep learning module, still deep learning has that gap over other algorithms as the output is again transferred to the initiator for improved measurement of the accuracy of the models. In many cases, the evaluation can be done through the f1-score where f1-score represents the harmonic score of precision and recall metrics which can normally provide the optimum results. Another metrics to compare these models are accuracy scores of the classifiers where the best accuracy score is selected as the best evaluation metrics for the different classifiers.

Table 4.3: Accuracy analysis of different models

Model	VADER	Naive Bayes	SVM	LSTM
Accuracy	68%	68%	71%	75%

For third classification, we used the SVM classification techniques that exhibited a model which provided the positive and negative class of the datasets. The data were sent inside the model and for the evaluation and data sets were divided into training and testing datasets. The SVM model was trained using training data and test data were used to evaluate the model created from training data. The following results were observed for SVM sentiment classifiers in the form of confusion matrix which showcase that the model gives 71% accurate results with 0.15 f1-score for 0 or negative class and 0.82 f1-score for positive class. From this result we can say that model performance for both positive and negative classes was excellent according to the f1-score of the both classes.

For the last classification, we used the deep learning algorithm. The LSTM model was used to deep learn the dataset by creating different hidden layers and 10% of the total dataset were used for testing whereas 90% of the dataset were used for training the LSTM model. A total of 120 LSTM hidden layers were introduced while there was 1 dense layer. With a batch size of 64, a total of 5 epochs were created to train the LSTM model. Normally the results are much more optimized while using any form of deep learning model. In our case, the LSTM model is used which follows the recurrent model and the outputs are adjusted such that they are feedback into the activation function giving the optimum output. Below figure explains the training of the data using all the parameters mentioned above up to 5 epochs.

After the creation of the LSTM model, the model was evaluated and the model bears a 75% accuracy while evaluating with the result. This showcases that deep learning method results in higher accuracy of the model other than unsupervised and supervised models.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

As the thesis aims for its objectives to be fulfilled, we clearly classified our data among two classes Positive and Negative using different kinds of algorithms. The specific objective was to classify our data set into Positive or Negative classes which is done by several classification algorithms like VADER, Naive-Bayes, SVM and LSTM. Before subjecting these processed data into the classifier, the data were first pre-processed using different techniques and features were extracted and then the data sets were divided into two phases:- Training and Testing phases. Furthermore, Visualization of the data along with the normality test is done which showcases that the datasets are normally distributed. Another specific objective was to perform analytics for different other e-commerce sectors based on the no. of positive and negative reviews. This is done after the sentiment analysis is done for each e-commerce sector and the data were visualized which showcased that Daraz leads the e-commerce sector in Nepal with very few negative sentiment about it regarding the products/services. Also different word clouds are created for the positive sentiments and negative sentiments.

Secondary data was collected from the popular e-commerce site of Nepal i.e. Daraz. and the classification model was created from the primary dataset obtained through survey. From the evaluation of the result of different algorithms, we can see that the deep learning model (LSTM) model exhibited 75% accuracy while the SVM model, which is a supervised learning model, exhibited 71% accuracy from our dataset. And then both the VADER and Naive-Bayes Classifier exhibited 68% accuracy. From the data observation, the data were classified into two classes 0 for Negative and 1 for Positive, Positive data's exhibit more accuracy than the negative data. Hence from this analogy, one can promote the ecommerce services in Nepal with different categories of the products and services and will get a positive response.

5.2 Recommendations

Till now, the datasets were collected from the questionnaire and some certain tests were performed. The data were visualized and the word cloud was generated after some pre-processing the reviews data.

Although we have sufficient data, it still is not adequate for machine learning algorithms to work at their best results. Generally, Deep learning techniques produce the most accuracy but here the SVM classifier has the highest accuracy. If the data sets are huge enough, more accurate calculations can be done. When the data sets are large, ML algorithms always produce nearly accurate results.

Another recommendation can be the sentiment analysis in Nepali language as there can be different datasets which contain nepali language to express the sentiments of a particular Nepali-language based sentence.

References

Bhatt, A. (2015). Amazon Review Classification and Sentiment Analysis. *International Journal of Computer Science and Information Technologies*, (5107-5110.).

Bonthu, H. (2021). *Rule-Based Sentiment Analysis in Python for Data Scientists*. Analytics Vidhya. Retrieved September 22, 2022, from https://www.analyticsvidhya.com/blog/2021/06/rule-based-sentiment-analysis-in-python/

Chakraborty, K., Bhattacharya, S., Bag, R., and Hassanien, A. A. (2018). Sentiment analysis on a set of movie reviews using deep learning techniques. *Social network analytics: Computational research methods and techniques*, 127. https://doi.org/10.1016/B978-0-12-815458-8.00007-4

Chen, T., Xu, R., He, Y., and Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN.

Garcia, A. L., Lucas, J. M. D., and Antonacci, M. (2020). A cloud-based framework for machine learning workloads and applications. *IEEE Access*, *1*(1). doi:10.1109/access.2020.296438

Graves, A. (2015). *Understanding LSTM Networks*. colah's blog. Retrieved September 21, 2022, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

guide. *Sentiment Analysis Guide*. MonkeyLearn. Retrieved September 22, 2022, from https://monkeylearn.com/sentiment-analysis/

Haque, T. U., Saber, N. N., and Shah, F. M. (2018). Sentiment analysis on large scale Amazon product reviews. *IEEE International Conference on Innovative Research and Development (ICIRD)*. doi:10.1109/icird.2018.8376299

Home. (2018,). YouTube. Retrieved September 20, 2022, from http://nlp.stanford.edu/IR-%20book/html/htmledition/dropping-common-terms-stop-words-1.html/

Jollie, I.T. (2002). Principal Component Analysis. *Springer*.

Ma, Y. (2020). *NLP: How does NLTK.Vader Calculate Sentiment? | by Ying Ma*. Medium. Retrieved September 22, 2022, from https://medium.com/@mystery0116/nlp-how-does-nltk-vader-calculate-sentiment-6c32d0f5046b

Matsuo, Y., and Ishizuka, M. (2004). Single Document using Word Co-occurrence Statistical Information. *International Journal on Artificial Intelligence Tools*, *13*, 157-169. https://doi.org/10.1142/S0218213004001466

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113. doi:10.1016/j.asej.2014.04.011

Menzli, A. (2022, July 21). *Tokenization in NLP: Types, Challenges, Examples, Tools - neptune.ai*. Neptune.ai. Retrieved September 22, 2022, from https://neptune.ai/blog/tokenization-in-nlp

Miao, Q., Li, Q., and Dai, R. (2009). AMAZING: A sentiment mining and retrieval system. *Expert Systems with Applications*, *36*(7192-71).

Ninghao, L., and Song, Q. (2018). Explaining RNN Predictions for Sentiment Classification. *Texas A&M University.*

Pandey, P., and Soni, N. (2019). Sentiment analysis on customer feedback data: Amazon product reviews. *Proceedings of IEEE international conference on machine learning, big data, cloud and parallel computing*, 320-322. 10.1109/COMIT-Con.2019.8862258.

P M, K. R., and D, J. S. (2021). Sentiment analysis, opinion mining and topic modelling of epics and novels using machine learning techniques. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.06.001

Rain, C. (2013). Sentiment Analysis in Amazon ReviewsUsing Probabilistic Machine Learning. *Swarthmore CollegeDepartment of Computer Science*.

Ray, S. (2017). *Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples*. Analytics Vidhya. Retrieved September 21, 2022, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Shaikh, T., and Deshpande, D. (2016). "Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews.

Sileyew, K. J. (2019). Research Design and Methodology. *Text Mining - Analysis, Programming and Application*. doi:10.5772/intechopen.85731

Singhal, G. (2020). *Importance of Text Pre-processing*. Pluralsight. Retrieved September 22, 2022, from https://www.pluralsight.com/guides/importance-of-text-pre-processing

Smola, A. J., and Scholkopf, B. (2002). Support Vector Machines and Kernel Algorithms.

Srinivas, V., Satyanarayana, C., Divakar, C., and Sirisha, K. P. (2021). Sentiment Analysis using Neural Network and LSTM. *IOP Conference Series Materials Science and Engineering*. http://dx.doi.org/10.1088/1757-899X/1074/1/012007

Suganya, E., and Vijayrani, S. (2019). Sentiment Analysis for Scraping of Product Reviews from Multiple Web Pages Using Machine Learning Algorithms. *Intelligent Systems Design and Applications*, 677–685. doi:10.1007/978-3-030-16660-1 66

tf–idf.. Wikipedia. Retrieved September 22, 2022, from https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Tripathy, A., Agrawal, A., and Rath, S. (2016). Classification of sentiment reviews using N-gram machine learning approach. *Expert Systems with Applications International Journal*.

Tyagi, E., and Sharma, A. K. (2017). Sentiment Analysis of Product Reviews using Support Vector Machine Learning Algorithm. *Indian Journal of Science and Technology*, *10*(35). DOI: 10.17485/ijst/2017/v10i35/118965.

Vaidhya, M., Shrestha, B., Sainju, B., Khaniya, K., and Shakya, A. (2017). Personality Traits Analysis from Facebook data. *21st International Computer Science and Engineering Conference(ICSEC)*. DOI:10.1109/ICSEC.2017.8443932

Wang, M. (2017). Text Mining for yelp dataset challenge. *University of California San Diego.*

Xu, S. (2018). Bayesian Naïve Bayes classifiers to classification. *Journal of Information Science*, *44*(1), 48-59.

Xu, S., Yang, S., and Lau, C.M. Keyword Extraction and Headline Generation Using Novel Word Features.

Zhao, H., Liu, Z., Yao, X., and Yang, Q. (2021). A machine learning-based sentiment anal-ysis of online product reviews with a novel term weighting and feature selection approach. *Information Processing and Management*, *58*(5). https://doi.org/10.1016/j.ipm.2021.102656.

Zucco, C., Calabrese, B., and Agapito, G. (2019). Sentiment analysis for mining texts and social networks data. *Methods and tools*. https://doi.org/10.1002/widm.1333

Appendix

Table A1: Showcasing Secondary dataset from Daraz

Categories	Rating	Reviews
Beauty	4	Thankyou :D so much for the Secretislandnepal Team Today I finally received the product … :P :P and I Love it . I am excited to use this sunscreen . The packing was good...
Beauty	4	Best products ever :E:E:S:S my skin become fair and glowing :S dark spots disappeared started 1 week ago i will buy again Thank u guska:S
Beauty	4	Very light and cool sunscreen. It worked really well for my skin. Doesn't feel sticky and heavy.
Beauty	4	This was my second time to purchase :P Al the product is good :L Thanks to daraz team too during pandemic days also they provide our choose item safely :D
Beauty	4	Quite decent product on it's price range. Not so premium, not so bad too. 50 - 50
Electronics	4	These earphones are worth it in this price segment.
Electronics	4	I want refund
Electronics	4	Bad quality
Electronics	4	Not value for money
Electronics	4	A little bit bulkier than expected but quality is good and it is worth it. If you’re looking for a wallet with lots of storage I will recommend this.
Fashion	4	looks almost black... although the product is pretty good.
Fashion	4	not good for money

Fashion	4	As a woman, I love using this wallet for men simply because of how minimal and spacious it is. Additionally, it was also way cheaper than the women’s wallet offered by wildhorn. Ladies, if you don’t mind looking a little masculine, this wallet is for you!
Fashion	4	The purse looks sturdy and well built. It has nice pockets for my many cards but unfortunately it didn't meet my requirements as after putting all those cards it was almost impossible to fold it without damaging my cards.