Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

Sentiment Detection From Social Media Analysis Using Supervised Machine Learning Approach

Author

Md.Asikul Islam

Roll No. 133064

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

Supervised by

Proff. Dr. Md. Rabiul Islam

Head

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

ACKNOWLEDGEMENT

At first I would like to pray to Almighty Allah ti give me the scope and enthusiasm for successful completion of my thesis work .

I would like to express my special appreciation and thanks to my supervisor **Proff. Dr. Md. Rabiul Islam**, Head, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, for his constant guidance,co-operation and every possible help thoughtful the work and preparation of this thesis. He has been a tremendous mentor for me. Again I would like to thank him for encouraging this research. His advice on both research as well as on my career have been priceless.

I am very grateful to all the respective teacher of Department of Computer Science Engineering ,Rajshahi University OF Engineering and Technology,Rajshahi for their valuable suggestion ,extending facilitation and inspiration from time to time.

I am also thankful to laboratory stuff of CSE Department for their co-operation and amiable behavior.

Finally , I would like to thank all those who help to create a nice environment in spite of many obstacle. I would like to thank our family and well-wishers for their encouragement and support.

Date: November, 2017 Md.Asikul Islam

RUET, Rajshahi

Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

With immense pleasure, it is hereby certified that the thesis ***Sentiment Detection From Social Media Analysis Using Supervised Machine Learning Approach****, is prepared by* ***Md.Asikul Islam****. Roll. 133064, has been carried out under my supervision. The thesis has been prepared in partial fulfillment of requirements for degree of Bachelor of Science in Computer Science and Engineering.*

Supervisor

_____

Proff. Dr. Md. Rabiul Islam
Head
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

External Examiner

_____

Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

ABSTRACT

The growing popularity of E-commerce, social medias, forums, blogs etc. created a new platform where anyone can discuss and exchange his/her views, ideas , suggestions and experience about any product or services.Social media is generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Tweeter sentiment analysis has been an effective and valuable technique for analysis people,s opinions. As the most widely used approach for tweet sentiment analysis, machine learning algorithms work well on the sentiment classification, just as they have been successfully applied for many other purposes.

In this thesis, we conduct a systematic and thorough empirical study on the machine learning algorithms for tweet sentiment analysis. It deals with identifying and classifying opinions or sentiments expressed in source text. Therefore, the main goal of the thesis is to investigate algorithms that can be applied for the opinion estimation. To that extend, data preprocessing and several experiments are conducted.We present a new feature and then, the classifier is trained and tested on a datasets with Logistic Regression and Support Vector Machine classifiers. In addition, frequency analysis on token, that make a feature for training data on the classifier is discussed.

Keywords: Sentiment analysis, Sentiment, Natural Language Processing,Text Mining, Feature selection,Supervised Machine Learning,Logistic Regression, Support Vector Machine.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Nowadays, online social media websites have been growing widely.Millions of people are using social network to express their emotions, opinion and share views about their daily lives.Social media provides an opportunity for businesses to connect with their customers. This can be use to know the feelings of users about their business, products, or topics of interest.

Twitter is one of the popular social media websites where users of twitter generate messages called tweets. Tweets expresses user's feelings on a particular things. Hence twitter can be taken as a valuable source of public sentiment .

Sentiment analysis is a social media analytics tool. Sentiment analysis is done by analyzing text and checking how many negative and positive keywords are present there . If the number of positive keywords is greater than negative keyword, it is considered positive content.If there are more negative keywords, it is called negative content. When there is a large ammount of data ,humans cannot handle the data .For this reason , Data Mining came in the front . We use here Machine learning approach to analysis Sentiment.Machine learning is a part of artificial intelligence and it contains many algorithms.Features are often captured and analyzed firstly and then algorithm's are used. The subjective tweet messages in Twitter are non-structured and length is confined to 140 character. So preprocessing and Feature selection also play important roles in tweet sentiment analysis. A variety of feature selection methods and pre-processing methods are often used. The performance of the different classifier are different. we have to find the best combination of learning algorithm, feature extracting method and pre-processing method . It is a challenge for applying machine learning to tweet sentiment classification.

Analysis of opinions plays an important role in all science areas (politics, economics, and social life). For example, in marketing, if the seller knows about the customer's satisfaction of particular product he/she may estimate demand on the product. The same for politicians, they will know whether people support them or not. Sentiment classification task is not new research area. However, the main focus of research was on the analysis of big documents (reviews).

In this thesis, We used the Twitter platform as a source of opinions. Our main interest is to investigation efficient algorithms that can be applied for classification purposes on tweets. Before using the algorithm,data preprocessing feature selection and data selection is done because algorithms accuracy can be affected by these.

To apply machine learning algorithms several steps should be performed: 1. Data collection. Tweets to be analyzed have to be retrieved from Twitter as well as the dataset for training purpose has to be obtained. 2. Preprocessing data. Tweets have to be pre-processed in order to remove the usernames, URLs, punctuation that do not contain any useful information. Moreover, words have to be lowercased. 3. Training process. Data that was extracted as training set is given to the classifier for learning. 4. Data classification. When training stage is complete the classifier can be used for analyzing polarity of tweets or reviews. At first, the classifier is fed with the testing dataset to check the accuracy of the algorithm then real data can be given to the classifier to extract sentiments from tweets. After machine learning algorithms are applied results are analyzed. Namely, accuracy of algorithms and their performance time are analyzed. Depending on results recommendations for improvement classification process is given.

For this purpose, two methods are studied. First is Logistic Regression and second is Support Vector Machine as classification methods. Therefore, the aim of the thesis is to perform experiments and investigate the performance of two different algorithms detecting positive and negative tweets/reviews. Furthermore, algorithm which gives better results has to be defined.

1.2 Motivation of work

There are many researchers working on sentiment analysis: some of them focus on increasing accuracy of classifiers, some of them work on difference between classifiers, and some of them research on new pre-processing methods that can significantly affect the performance of classifiers. However, most research works are on general normal texts, such as pro duct reviews, comments on news, movie reviews, instead on short and messy texts, like tweets.

Tweeter sentiment analysis is going to very important in different field. Brand Monitoring: One of the most well documented uses of Sentiment Analysis is to get a full 360 view of how your brand, product, or company is viewed by your customers and stakeholders. Widely available media, like product reviews and social, can reveal key insights about what your business is doing right or wrong. Companies can also use sentiment analysis to measure the impact of a new product, ad campaign, or consumer's response to recent company news on social media.

Sentiment Analysis for Customer Service Customer service agents often use sentiment analysis to automatically sort incoming user email into "urgent" or "not urgent" buckets based on the sentiment of the email, proactively identifying frustrated users. The agent then directs their time toward resolving the users with the most urgent needs first. As customer service becomes more and more automated through Machine Learning, understanding the sentiment of a given case becomes increasingly important.

Sentiment Analysis for Market Research and Analysis Sentiment analysis is used in business intelligence to understand the subjective reasons why consumers are or are not responding to something (e.x. why are consumers buying a product? What do they think of the user experience? Did customer service support meet their expectations?). Sentiment analysis can also be used in the areas of political science, sociology, and psychology to analyze trends, ideological bias, opinions, gauge reactions, etc. [1]

Recently, there has been some research on tweet sentiment analysis from the machine learning approach. A variety of machine learning algorithms have been studied for this purpose. there is a lack of a systematic study on the machine learning methods for tweet sentiment analysis. For this topic, we will choose some different pre-processing methods and evaluate in different evaluation methods. Because of a variety of combinations, we will give a systematic study and experiment to demonstrate the performance of various machine learning algorithms.

1.3 Objectives of the work

The aim of this thesis work is to investigate and discover efficient algorithms that can be used for sentiment analysis as well as to provide improvements for existing solutions. To that extend, following steps should be done:

- Investigate feature selection methods for text classification.
- Investigate machine learning algorithms that can be applied for classification problem.

- Analyze accuracy of these algorithms with respect to different datasets
- Analyze computational time of the algorithms and computational resources that particular algorithm requires.
- Compare results of applied techniques.
- Based on obtained result, propose a set of improvements.

1.4 Applications

- Commerce: Companies can make use of this research for gathering public opinion related to their brand and products. From the company's perspective the survey of target audience is imperative for making out the ratings of their products. Hence Twitter can serve as a good platform for data collection and analysis to determine customer satisfaction.
- Politics: Majority of tweets on Twitter are related to politics. Due to Twitter's widespread use, many politicians are also aiming to connect to people through it. People post their support or disagreement towards government policies, actions, elections, debates etc. Hence analyzing data from it can help is in determining public view.
- Sports Events: Sports involve many events, championships, gatherings and some controversies too. Many people are enthusiastic sports followers and follow their favorite players present on Twitter. These people frequently tweet about different sports related events. We can use the data to gather public view of a player's action, team's performance, official decisions etc.

1.5 Thesis outline

The rest of the thesis has four more chapters and organized as follows:
Chapter 2 is dedicated for the methodology of the Sentiment analysis that have been applied in other research paper on twitter sentiment analysis.
Chapter 3 is dedicated for the Background study of the Sentiment analysis .Different Method of feature selection and machine learning algorithm for twitter sentiment analysis have been described in this section.

Chapter 4 is dedicated for the methodology of the system in details. Flowchart, Algorithm are described section by section. In the subsection, Feature extraction, combination of feature methods are described. Sample description about Logistic and SVM classifier is given.
Chapter 5 depicts the experimental evaluation of the system. Dataset, experimental setup are described in details in this chapter. Implementation procedure and cross-validation are also discussed in this chapter.
Chapter 6 is dedicated for the evaluation matrix, result and the performance analysis. It contains Comparison,and Error Analysis.
Chapter 7 represents the summary of this research work and highlights the overall contribution. A direction also given on the future scope of math word problem solving.
Appendix A presents the system installation details.

1.6 Conclusion

In this chapter ,we discuss about the short summmary of the whole research,objectives,motivation and outline.the target is to find the best combination of feature selection method and machine learning algorithm.there are various method for sentiment analysis.we focus on less computation,time efficiency and best accurate method and implement this approach on tweet data

Chapter 2

Literature Review

2.1 Introduction

Sentiment analysis is not a new task, it has been studied since 90s. However, in 2000s Sentiment Analysis attracted the interest of scientists due to its significance in different scientific areas, also Sentiment Analysis had a many unstudied research questions [2]. Sentiment analysis is a developing area that is used in the field of humans interest and especially organizations because Sentiment Analysis can be used for decision making process. Sentiment analysis deals with processing of opinionated text in order to extract and categorize opinions from certain document. The polarity of sentiment usually expressed in terms of positive or negative opinion.

Supervised machine learning methods assume that training data have the label and they are used for the learning process. Then the output cn be estimated from the input dataset.

2.2 Related Works

David Zimbra ; M. Ghiassi ,Sean Lee,"Brand-related Twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks" [3].

The feature engineering produced a final tweet feature representation consisting of only seven dimensions, with greater feature density on a Starbucks brand-related Twitter data set. the three-class tweet sentiment classification accuracies with SVM classifier exceeds 78 percents .

There proposed feature engineering method conduct a pairwise t-test and the experiment results is , n=2179 for three class classification.For SVM , multiple binary classifier were de-

veloped to perform one vs. all sentiment classifier. Overall Sentiment classification accuracy by SVM is 78.33

Balakrishnan Gokulakrishnan,AShenan Perera, ,"Opinion mining and sentiment analysis on a twitter data stream" [4] discusses an approach where a publicized stream of tweets from the Twitter microblogging site are preprocessed and classified based on their emotional content as positive, negative and irrelevant; and analyses the performance of various classifying algorithms based on their precision and recall in such cases.They use following step for preprocessing the dataset: A. Replacing Emoticons B. Uppercase Identification C. Lower Casing D. Url Extraction E. Detection of Pointers F. Identification of punctuations G. removal Of Stop Words H. Removal of Query Term I.Compression Of Word G. Removing Skewness in Dataset

They use different machine learning classifier .Among them the Performance of some algorithm are described follow:

Bayesian Logistic Regression :

Avg. Accuracy-75.03% , Max Accuracy-76.62% ; Avg.F-0.75

Support Vector Machine:

Avg. Accuracy-72.70% ;Max Accuracy-72.70% ; Avg.F-0.727

In this Paper we focus on Supervised Machine learning approach for Sentiment Analysis.Algorithms are design such a way that computers can learn from different task. Feature are extract from dataset and algorithm learn from those feature. Usually,machine learning algorithms work well on inferring information about the properties of sets of data. In this paper ,For the task of sentiment analysis, we want to find a feature that is most relevant and add this with different feature extraction technique.The success of machine learning also relies on selection and extraction of sentiment features,which are especially from natural language processing (NLP) techniques. Different machine learning techniques might be used to handle different sentiment classification problems. From the previous work, many researchers choose machine learning approaches to deal with their tasks of sentiment analysis. They experimented different machine learning algorithms, such as Naive Bayes classifier, Support Vector Machine (SVM) classifier, Maximum Entropy classifier and so on. Because of using different pre-processing methods and different training data, these classifiers can give different performance.

Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1.12 (2009). [5] discuss an approach of auto-

matically classifying the sentiment of Twitter messages.These messages are classified as either positive or negative with respect to a query term.They present the results of machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision.The machine learning classifiers are Naive Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM). The feature extractors are unigrams, bigrams, unigrams and bigrams, and unigrams with part of speech tags. They build a framework that treats classifiers and feature extractors as two distinct components.

Using of unigram feature extractor , they report 82.2% accuracy for Support Vector Machine.Using of bigram feature extractor , they report 78.8% accuracy for Support Vector Machine.Using both unigram and bigram feature extractor , they report 81.6% accuracy for Support Vector Machine.

Chapter 3

Background

3.1 Introduction

In the following sections background information for sentiment analysis is presented for this paper. First of all,The definition of Sentiment analysis is discussed. Secondly, Different type of analyzing sentiment data is discussed. Thirdly,Machine learning and text mining is defined.Fourthly, Methods for s Sentiment Analysis is discussed.Fifthly,Discussed about different Feature extraction technique.Then,Different Supervised Method of sentiment analysis is discussed.

3.2 Sentiment analysis

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, ”beyond polarity” sentiment classification looks, for instance, at emotional states such as ”angry”, ”sad”, and ”happy”. According to Bing Liu in [2]: ”sentiment analysis, also called opinion mining, is the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.” In the following sections sentiment analysis will be explained. Sentiment analysis is the computational study of people’s opinions, appraisals, and emotions toward entities, events and their attributes (Liu, 2010 pg.1).

The section will start with a brief introduction of the history of sentiment analysis and where

it originated. Following this, an overview will be given about how sentiment analysis works.

3.2.1 Origin of sentiment analysis

Pang and Lee [6]give three reasons why the interest in sentiment analysis is flourishing. The first reason is the rise of machine learning methods in the field of information retrieval and natural language processing. The second reason is that via the Internet many review sites emerged. This resulted in a wide availability of datasets that can be used for machine learning algorithms. Finally, the area offers interesting commercial and intelligence applications.

3.2.2 Basics of sentiment analysis

To better understand the principles of sentiment analysis, [2] uses six steps to describe sentiment analysis. These six steps will be explained in the following section using an 18 example tweet. The example tweet is: I think the HTC One M8's Camera gets a lot of hate. Using it for the last two days I don't think its as bad as what people think it is (15:17 - 28 mrt. 2014)13. The first step that is described by Liu is the entity extraction and categorization. In the case of the example tweets this means that the sentiment analysis tool should extract the words HTC One. With the categorization is meant that synonyms that are similar to HTC should also be extracted and categorized together, or put into clusters. The second step of sentiment analysis is aspect extraction and categorization. Here all the aspects that are connected to HTC One should be extracted from the text and be connected to the entity. In this case this would mean that the word Camera should be extracted. So in the second step every aspect that tells something about the entity should be extracted. Other examples that say something about HTC One could be 'Battery life', 'Screen', and 'Picture quality'. The third step is the opinion holder extraction and categorization. This is recognizing the opinion of the writer of the text, also referred as the sentiment of the text. Recognizing an opinion or sentiment is done by comparing words to a lexicon with words that have a known sentimental value (Raijmakers, 2013). Words in a lexicon can have positive sentiment such as good and beautiful, or they can have negative sentiment such as useless and bad. In the case of the example tweet the word that gives sentimental value is 'bad' and the word 'hate'. Even though, bad is the word with sentimental value the author means the opposite of bad. The author means 'I don't think its as bad as what people think it is'. This provides sentiment analysis tool with an extra challenge to provide tweets with

an accurate sentimental value.

The fourth step is time extraction and standardization. The sentiment analysis tool should find the time and date the message is posted. In the case of the example this is on 15:17, March 18th, 2014. The fifth step is aspect sentiment classification. Here the sentiment of the entire text message is determined, which could either be positive, negative, or neutral. In the case of the example tweet this should give the tweet the classification as positive. The sixth step is to create an overview of all the previous described steps. This way the user of the sentiment analysis tool can see that the tweet is positive tweet about the camera of the HTC M8.

3.3 Type Of Sentiment data Analysis

There are different methods to analyze the 'sentiment data'. Let us take a look at each of them here.We discuss about them follwing: [7]

3.3.1 Document-level of sentiment analysis:

Opinions describe people's sentiments, appraisals or feelings towards an entity or an event. Many blogs or forums allow people to express their opinion in the form of reviews and comments. When opinions are expressed in the form of reviews, instead of a simple 'Yes' or 'No', identifying the actual emotions would need a subjective analysis of the words used in the review. [8] In document-level of sentiment analysis, each document focuses on a single entity or event and contains opinion from a single opinion holder [9] .The opinion here are can be classified in to two simple classes: Positive or negative (probably neutral). For example: A product review: "I bought a new phone few days ago. It is a nice phone, though it is a little big. The touch screen is good. The voice clarity is better. I simply love the phone" . Considering the words or phrases used in the review (nice, good, better, love), the subjective opinion is said to be positive. The objective opinions are measured using the star or poll system, where 4 or 5 stars are positive and 1 or 2 stars are negative.

3.3.2 Sentence-level of sentiment analysis:

To have more refined view of different opinions expressed in the document about the entities, we should move to the sentence level. a sentiment score is generated for each sentence in this

analysis.This level of sentiment analysis – filters out those sentences which contain no opinion and – determines whether the opinion on the entity is positive or negative.

3.3.3 Aspect based sentiment analysis

Document level and sentence level sentiment analysis works well when they refer to a single entity. However, in many cases people talk about entities that have many aspects or attributes. They will also have different opinions about different aspects. It often happens in product review and discussion forums. For example: ”I am a Nokia phone lover. I like the look of the phone. The screen is big and clear. The camera is fantastic. But, there are few downsides too; the battery life is not up-to the mark and access to Whatsapp is difficult.” Categorizing the positive and negatives of this review hides the valuable information about the product. Therefore, the Aspect –based sentiment analysis focuses on [10] the recognition of all sentiment expressions within a given document and the aspects to which the opinions refer.

3.3.4 Comparative sentiment analysis

In many cases, users express their opinions by comparing it with a similar product or brand. Therefore, the goal here is to identify sentences that contain comparative opinions. For example: ”I drove the Honda Civic, it does not handle better that the Skoda Superb”

3.3.5 Sentiment lexicon acquisition:

This sentiment analysis method uses a list of words and expressions used to express people's subjective feelings and sentiment or opinions. It not only uses certain words, but also phrases and idioms. In the other types of sentiment analysis, we have seen what positive and negative words are. Let us take an example: ”Car X is better than car Y.” This sentence does not express an opinion that any of the two cars is good or bad. Therefore, these types of sentences/documents are furthered analyzed using 3 approaches: Manual approach, dictionary based approach and corpus-based approach [7].

Manual Approach: This is not feasible as it is time consuming.

Dictionary based approach: This approach uses ‘Word Net’ to find suitable words of the sentiment word to carry out the analysis.

Corpus-based approach: This is used to create a domain-specific sentiment lexicon to carry out the analysis.

3.4 Machine learning and Text Mining

Text mining is the process of examining large collections of written resources to generate new information and to transform the unstructured text into structured data for use in further analysis. The first stage of text or data mining is to retrieve information. This might require using a search engine to identify a corpus of texts that are already digitized or it might necessitate digitalization of physical texts in publications or manuscripts. This corpus will need to be brought together in a useful format . The second stage is the mark-up of text to identify meaning. In most cases this will involve adding meta-data about the text into a database , while in others it might involve keying in all person names or locations mentioned in the text . This process allows search engines to extract information and identify relationships between texts based upon the preconceptions of those creating the meta-data [11].

The final stage is to text mine the text(s) using various tools. The purpose is to find associations among pieces of information that draw out meaning and enable researchers to discover new information which might otherwise be difficult to discover.

3.5 Methods for Sentiment Analysis

Sentiment Analysis can be performed in three ways: 1) Sentiment Analysis based on Supervised Machine learning technique, 2) Sentiment Analysis by using Lexicon based Technique and 3) Sentiment Analysis By combining the above two approaches. [12]

3.5.1 Supervised Machine learning based techniques

In Supervised Machine learning techniques, two types of data sets are required: training dataset and test data set. An automatic classifier learns the classification factors of the document from the training set and the accuracy in classification can be evaluated using the test set. Various machine learning algorithms are available that can be used very well to classify the documents. The machine learning algorithms like Support Vector Machine (SVM), Naive Bayes (NB) and maximum entropy (ME) are used successfully in many research and they performed well in the

sentiment classification. The first step in Supervised Machine learning technique is to collect the training set and then select the appropriate classifier. Once the classifier is selected, the classifier gets trained using the collected training set. The key step in the Supervised Machine learning technique is feature selection. The classifier selection and feature selection determines the classification performance. The most common techniques used for feature selection are: 1) Opinion words and phrase: By considering adjectives and adverb most of the opinion words can be extracted from the document, but sometimes nouns or verbs may also express opinion. For instance, good, fantastic, amazing, bad and boring are all adjective or adverb which express emotions while rubbish is a noun but it express a sentiment similarly hate and like are verb but it express opinion. Once opinions are collected, its polarity can be calculated using statistical-based or lexiconbased techniques. 2) Terms and their frequency: uni-grams or n-grams with their frequency of occurrence are considered as features. This technique is used in many studies and achieved good result. Pang et al. [13] used uni-grams on movie review dataset and Dave et al. [14] used bigrams and tri-grams on product review dataset. Both studies reported better result on polarity determination. 3) Part of speech (POS) information: In this approach, POS tag of words is used in determining the feature. In POS tagging, it tag each word by considering its position in the grammatical context. Prabowo and Thelwall(2009) [15] used this approach in their studies and they constructed feature set easily by identifying adjectives and adverbs. 4) Negations: Negation word reverses the meaning, so it very important factor in polarity calculation [6]. Pang et al. (2002) [13] used three supervised machine learning techniques to classify the text: SVM, Naïve Bayes, and Maximum Entropy. They compared the efficiency of these three classifiers with different feature selection method such as uni-gram, n-gram, combining uni-gram and bi-gram and by combining uni-gram and Pos tagging. They reached a conclusion that if the feature set is small then it is better to consider feature presence than feature frequency. While Naïve Bayes performed well on small feature set, The SVM performed well on large feature space. Maximum Entropy gave better result than Naïve Bayes when experimented with large feature space. Abbasi et al. [16] evaluated the utility of stylistic and syntactic features for sentiment classification in English and Arabic language. Structural and lexical attributes contribute to stylistic feature while manual, semiautomatic, or automatic annotation techniques are used for syntactic features. A new hybridized genetic algorithm, the entropy weighted genetic algorithm (EWGA) is introduced to improve feature selection process that incorporates the information-gain heuristic. They achieved an accuracy of over 95classifier

on movie review data set. The Stylistic feature enhanced performance in all test sets.

3.5.2 Lexicon Based Method

Lexicon Based Method is an Unsupervised Learning approach since it does not require prior training data sets. [17] It is a semantic orientation approach to opinion mining in which sentiment polarity of features present in the given document are determined by comparing these features with semantic lexicons. Semantic lexicon contains lists of words whose sentiment orientation is determined already. It classifies the document by aggregating the sentiment orientation of all opinion words present in the document, documents with more positive word lexicons is classified as positive document and the documents with more negative word lexicons is classified as negative document. The key steps of lexicon based sentiment analysis are the following [18]: 1. Preprocessing: This step clean the document by removing HTML tags and noisy characters present in the document, by correcting spelling mistakes, grammar mistakes, punctuation errors and incorrect capitalization and replacing non-dictionary words such as abbreviations or acronyms of common terms with their actual term. 2. Feature Selection: This step Extract the feature present in the document by using techniques like POS tagging. 3. Sentiment score calculation: Initialize s with 0. For each extracted sentiment word, check whether it is present in the sentiment dictionary, If present with negative polarity, w then s = s-w or If present with positive polarity, w then s = s + w. 4. Sentiment Classification: If s is below a particular threshold value then classifying the document as negative otherwise classify it as positive.

3.5.3 Hybrid Techniques

Some researchers combined the supervised machine learning and lexicon based approaches together to improve sentiment classification performance. Fang et al. [19]adopted entirely different approach. They considered both general purpose lexicon and domain specific lexicon for determining polarity orientation of sentiment words and feed these lexicons into supervised learning algorithm, SVM. They found that general purpose lexicon performed very poor while domain specific lexicon performed very well. The system classified the sentiment in two steps: First the classifier is trained to predict the aspects and In Next the classifier is trained to predict the sentiments related to the aspects collected in step1.Their system yielded around 66.8Mudi-

nas et al. [20] combined lexicon based and learningbased approaches to develop a concept-level sentiment analysis system, pSenti. It utilized advantages of both the approaches and attained stability and readability from semantic lexicon and high accuracy from a powerful supervised learning algorithm. They extracted sentiment words and considered it as features in machine learning algorithm. This hybrid approach pSenti achieved an accuracy of 82.30%. Zhang et al. [21]carried out entity level sentiment analysis. They utilized both the supervised learning techniques and lexicon based techniques. By lexicon based method they extracted sentiment words. By using Chi-square test on the extracted seeds additional seeds are discovered. Sentiment polarities of newly discovered seed are determined through a classifier, which is being already, trained using initial seeds. There is no manual task in this proposed system and it achieved around 85.4% of accuracy [12].

3.6 Feature Extraction

If we want to use text in machine learning algorithms, we'll have to convert them to a numerical representation. One of the methods is called bag-of-words approach. The bag of words model ignores grammar and order of words. Once we have a corpus (text data) then rst, a list of vocabulary is created based on the entire corpus. Then each document or data entry is represented as numerical vectors based on the vocabulary built from the corpora

3.6.1 Count Vectorizer

With count vectorizer, we merely count the appearance of the words in each text. For example, let's say we have 3 documents in a corpus: "I love dogs", "I hate dogs and knitting", "Knitting is my hobby and my passion". If we build vocabulary from these three sentences and represent each document as count vectors, it will look like below pictures [22].

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

Figure 3.1: Diagram of Count Vectorizer Method

3.6.2 The Bag of Words representation

We need a way to represent text data for machine learning algorithm and the bag-of-words model helps us to achieve that task. The bag-of-words model is simple to understand and implement. It is a way of extracting features from the text for use in machine learning algorithms.

Source In this approach, we use the tokenized words for each observation and find out the frequency of each token. Let's take an example to understand this concept in depth.

"It was the best of times" "It was the worst of times" "It was the age of wisdom" "It was the age of foolishness"

We treat each sentence as a separate document and we make a list of all words from all the four documents excluding the punctuation. We get,

'It', 'was', 'the', 'best', 'of', 'times', 'worst', 'age', 'wisdom', 'foolishness'

The next step is the create vectors. Vectors convert text that can be used by the machine learning algorithm.

We take the first document:"It was the best of times" and we check the frequency of words from the 10unique words

"it" = 1 "was" = 1 "the" = 1 "best" = 1 "of" = 1 "times" = 1 "worst" = 0 "age" = 0 "wisdom" = 0 "foolishness:" = 0

Rest of the documents will be: "It was the best of times" = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0] "It was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0] "It was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0] "It was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]

In this approach, each word or token is called a "gram". Creating a vocabulary of two-word pairs is called a bigram model.

For example, the bigrams in the first document : ”It was the best of times” are as follows: ”it was” ”was the” ”the best” ”best of” ”of times”

The process of converting NLP text into numbers is called vectorization in ML. Different ways to convert text into vectors are:

Counting the number of times each word appears in a document. Calculating the frequency that each word appears in a document out of all the words in the document [23].

3.6.3 N-Gram Modeling:Unigram,Bigram,Trigram

n-gram is a continuous sequence of n items from a given sequence of text or speech”. In other words, n-grams are simply all combinations of adjacent words or letters of length n that you can find in your source text. Below picture represents well how n-grams are constructed out of source text [22].

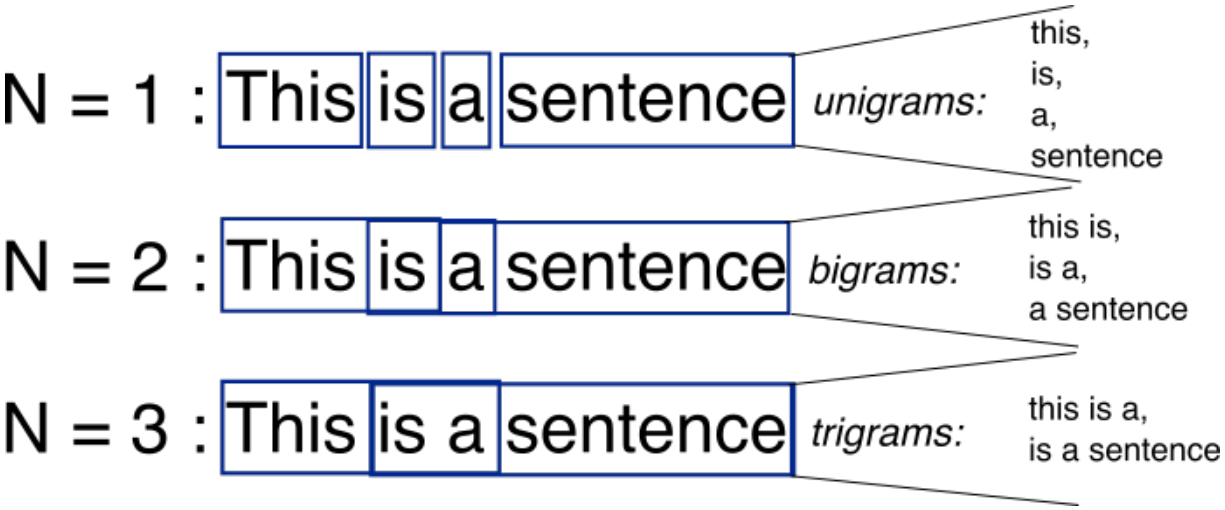


Figure 3.2: Diagram of Top tokens in tweet Dataset

3.6.4 TF-IDF

TFIDF is another way to convert textual data to numeric form, and is short for Term Frequency-Inverse Document Frequency. The vector value it yields is the product of these two terms; TF and IDF.

Let’s first look at Term Frequency. We have already looked at term frequency with count vectorizer, but this time, we need one more step to calculate the relative frequency. Let’s say we have two documents in our corpus as below.

1.I love dogs 2.I hate dogs and knitting Relative term frequency is calculated for each term within each document as below.

$$TF(t,d) = \frac{number\ of\ times\ term(t)\ appears\ in\ document(d)}{total\ number\ of\ terms\ in\ document(d)}$$

For example, if we calculate relative term frequency for 'I' in both document 1 and document 2, it will be as below.

$$IDF('I',D) = \log\left(\frac{2}{2}\right) = 0$$

Next, we need to get Inverse Document Frequency, which measures how important a word is to differentiate each document by following the calculation as below.

$$IDF(t,D) = \log\left(\frac{total\ number\ of\ documents(D)}{number\ of\ documents\ with\ the\ term(t)\ in\ it}\right)$$

If we calculate inverse document frequency for 'I',

$$IDF('I',D) = \log\left(\frac{2}{2}\right) = 0$$

Once we have the values for TF and IDF, now we can calculate TFIDF as below.

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D)$$

Following the case of our example, TFIDF for the term 'I' in both documents will be as below.

$$TFIDF('I',d1,D) = TF('I',d1) \cdot IDF('I',D) = 0.33 \times 0 = 0$$

$$TFIDF('I',d2,D) = TF('I',d2) \cdot IDF('I',D) = 0.2 \times 0 = 0$$

As you can see, the term 'I' appeared equally in both documents, and the TFIDF score is 0, which means the term is not really informative in differentiating documents. The rest is same as count vectorizer, TFIDF vectorizer will calculate these scores for terms in documents, and convert textual data into the numeric form. [22]

3.7 Different Supervised Classificaton Methods

3.7.1 SVM

A Support Vector Machine models the situation by creating a feature space, which is a finite-dimensional vector space, each dimension of which represents a "feature" of a particular object. In the context of spam or document classification, each "feature" is the prevalence or importance of a particular word.

The goal of the SVM is to train a model that assigns new unseen objects into a particular category. It achieves this by creating a linear partition of the feature space into two categories. Based on the features in the new unseen objects (e.g. documents/emails), it places an object "above" or "below" the separation plane, leading to a categorisation (e.g. spam or non-spam). This makes it an example of a non-probabilistic linear classifier. It is non-probabilistic, because the features in the new objects fully determine its location in feature space and there is no stochastic element involved.

However, much of the benefit of SVMs comes from the fact that they are not restricted to being linear classifiers. Utilising a technique known as the kernel trick they can become much more flexible by introducing various types of non-linear decision boundaries.

3.7.2 Logistic Regression

Logistic regression is generally used where the dependent variable is Binary or Dichotomous. That means the dependent variable can take only two possible values such as "Yes or No", "Default or No Default", "Living or Dead", "Responder or Non Responder", "Yes or No" etc. Independent factors or variables can be categorical or numerical variables.

Please note that even though logistic (logit) regression is frequently used for binary variables (2 classes), it can be used for categorical dependent variables with more than 2 classes. In this case it's called Multinomial Logistic Regression.

Here we will focus on Logistic Regression with binary dependent variables as it is most commonly used.

Applications of Logistic Regression-

Logistic regression is used for prediction of output which is binary, as stated above. For example, if a credit card company is going to build a model to decide whether to issue a credit card to a customer or not, it will model for whether the customer is going to “Default” or ”Not Default” on this credit card. This is called “Default Propensity Modeling” in banking lingo.

Similarly an ecommerce company that is sending out costly advertisement / promotional offer mails to customers, will like to know whether a particular customer is likely to respond to the offer or not. In Other words, whether a customer will be ”Responder” or ”Non Responder”. This is called ”Propensity to Respond Modeling”

Using insights generated from the logistic regression output, companies may optimize their business strategies to achieve their business goals such as minimize expenses or losses, maximize return on investment (ROI) in marketing campaigns etc.

Chapter 4

Methodology

4.1 Introduction

Nowadays, we we have a tendency to accessing the internet everyday .we cannot imagine our life without internet. Everyone uses Internet for different purposes, i.e. for searching some information or posting something. Information can be easily published by users in the blogs, forums, social networks, feedbacks can be left on the particular web pages. There are a lot of web-sites that provide business and product reviews. For example, TripAdvisor is a website that provides dozens of opinionated information about hotels, restaurants, flights, places where to go, which is very helpful for travelers. Twitter is another way of sharing views. Information from such sources is used not only by customers, but it is also vital for different organizations. The availability of data make the need to create of an automated system for searching and classifying opinions.

4.2 Problem statement

Sentiment analysis on Twitter messages is conduct by identifying positive and negative ones. For Sentiment Analysis, selecting data sets on which the experiments are conducted, is also important. For sentiment analysis over Twitter, a number of benchmark datasets have been released in the last few years, and they are available online. Furthermore, we prefer to select the datasets that have been widely used in the Twitter sentiment analysis experiments in literatures [5]. Classification is performed on tweets, where each tweet is labeled as positive or negative according to the opinion expressed in it. Before building the classifier, training data have to

be collected and preprocessed in order to discard irrelevant information from the training set. Moreover, preprocessing should be performed to reduce the size of training dataset, which in turn may lead to speeding up the training process. The next important step that has to be done until training the model is feature selection. Feature selection allows to build a set of unique terms (features) across the corpus by excluding ambiguous terms. After the classification model is created and tested, parameters of accuracy, precision and recall as well as computation time have to be estimated. Furthermore, comparison of the applied algorithms has to be provided according to the classification results.

4.3 Existing System

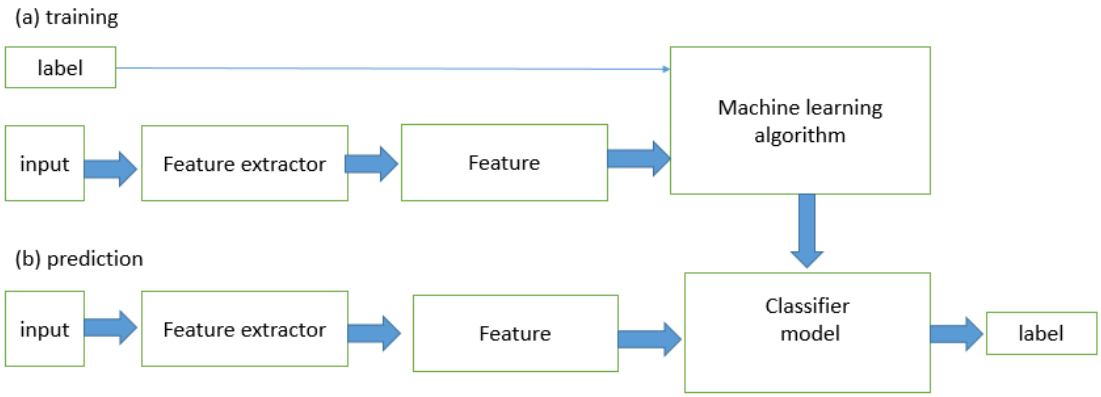


Figure 4.1: The architectural overview for existing system

4.4 Proposed framework

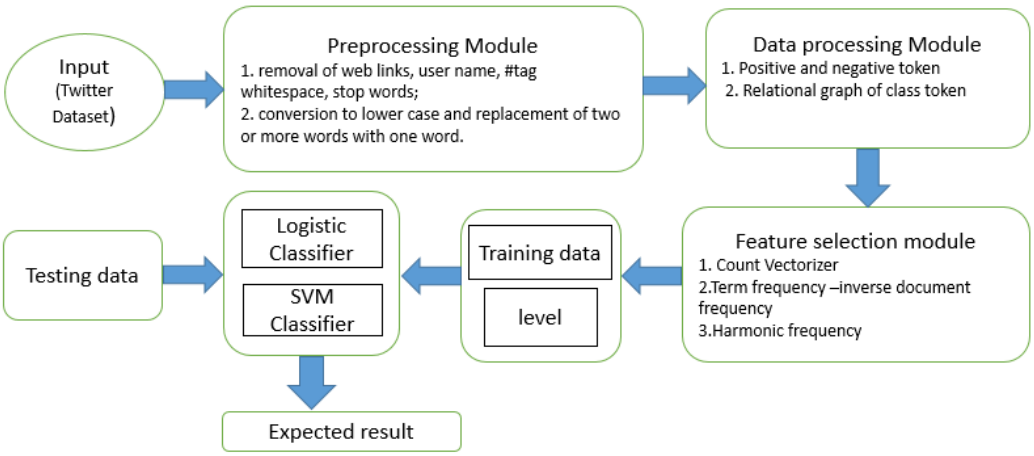


Figure 4.2: The architectural overview for proposed system

4.5 Procedure of Sentiment Analysis

For Sentiment analysis , the most important thing is feature selection. The more accurately feature can be select , the system will be more perfect. After the selection of feature , a classification algorithm is applied and training data trained the classifier for analysis the unknown data.

4.5.1 Feature extraction

In sentiment analysis ,collected text isn't in a form understandable by a computer.For thid reason , feature need to be extracted from text data and these feature help to learn a classifier algorithm. In the background ,we have disscuss different feature extraction method like count vectorizer,term frequency-inverse document frequency . Comparing the existing method and proposed method , we can see that we add a feature and combine the feature with existing feature extraction method.

Adding a new feature follow the following technique:

we have firstly generate frequency of all token and plot the relationship of positive token frequency and negative token frequency .the plot can not find any pattern.

Then, we have generate harmonic frequency of all token and plot the relationship of positive token frequency and negative token frequency .the plot can not find any pattern.

Finally, we have generate normalized cumulative frequency of all token and plot the relationship of positive token frequency and negative token frequency .CDF (Cumulative Distribution Function) value of both pos-rate and pos-freq-pct. CDF can be explained as ”distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x”. By calculating CDF value, we can see where the value of either pos-rate or pos-freq-pct lies in the distribution in terms of cumulative manner. harmonic mean of rate CDF and frequency CDF has created an interesting pattern on the plot. If a data point is near to the upper left corner, it is more positive, and if it is closer to the bottom right corner, it is more negative.

So ,the normalized cumulative Distribution harmonic frequency is most relevant feature for this dataset.

4.5.2 Classification algorithms

We identify problem as classification problem when independent variables are continuous in nature and dependent variable is in categorical form i.e. in classes like positive class and negative class. The real life example of classification example would be, to categorize the mail as spam or not spam, to categorize the tumor as malignant or benign and to categorize the transaction as fraudulent or genuine. All these problem's answers are in categorical form i.e. Yes or No. and that is why they are two class classification problems. There has many classification algorithm and we have applied logistic Regression and Support Vector Machine for the dataset. Here is a description of those algorithm.

4.6 Logistic Regression

Logistic regression is a technique borrowed by machine learning from the field of statistics.It is the go-to method for binary classification problems (problems with two class values). It is one of the basic and popular algorithm to solve a classification problem. It is named as ”Logistic Regression”, because it's underlying technique is quite the same as Linear Regression. The term ”Logistic” is taken from the Logit function that is used in this method of classification. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.When we want to look at a dependence structure, with a dependent variable and a set of explanatory variables (one or more), we can use the logistic regression framework. Here a figure: [24]

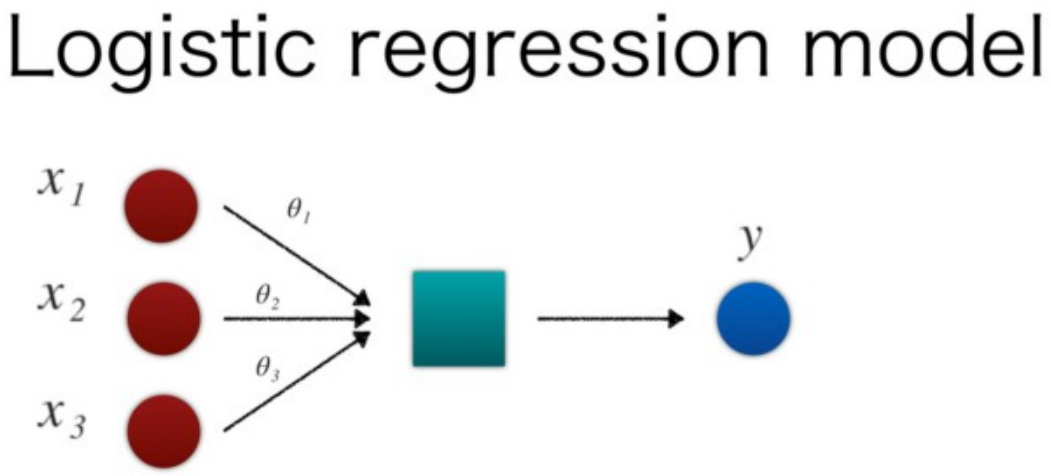


Figure 4.3: Diagram of logistic regression model

4.6.1 Mathematical Foundation

logistic function:

An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is a Sigmoid function, which takes any real value between zero and one. It is defined as $g(z) = \frac{1}{1+exp(-z)}$ where $z = \alpha 1x + \alpha 2$

if we plot it, the graph will be a curve [25]:

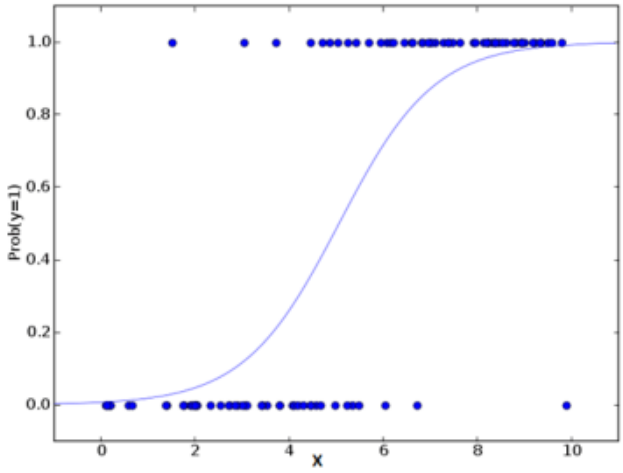


Figure 4.4: Diagram of logistic function

The logistic regression hypothesis is then defined as:

$$h_\theta(x) = g(\theta^T x) \tag{4.1}$$

Logistic regressions are usually fit by maximum likelihood. The cost function we want to minize is the opposite of the log-likelihood function:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y_i log(h_\theta(x_i) - (1-y_i) log(1 - h_\theta(x_i))] \tag{4.2}$$

This imply to solve the following equation:

$$\frac{dJ(\theta)}{d\theta} = \frac{1}{m}\sum_{i=1}^{N} x_i(h_\theta(x_i) - y_i) = 0 \tag{4.3}$$

4.6.2 Working Procedure

Decision Boundary:

Decision boundary helps to differentiate probabilities into positive class and negative class.

Linear Decision Boundary [25]:

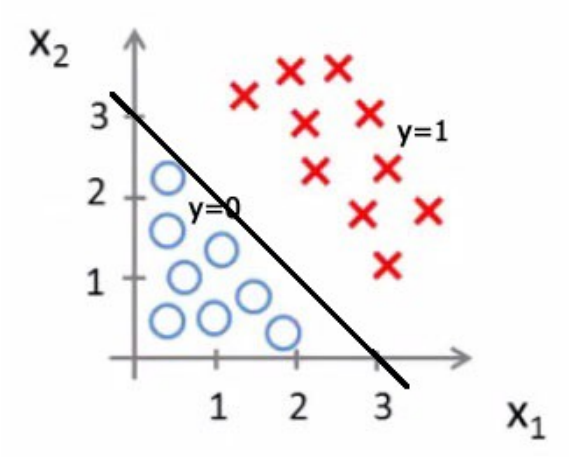


Figure 4.5: Linear Decision Boundary

Non Linear Decision Boundary [25]:

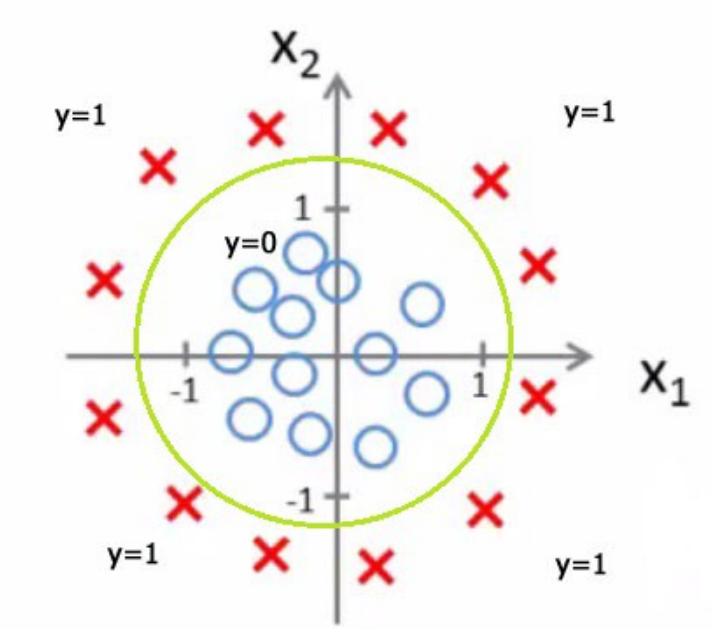


Figure 4.6: Non Linear Decision Boundary

4.7 SVM

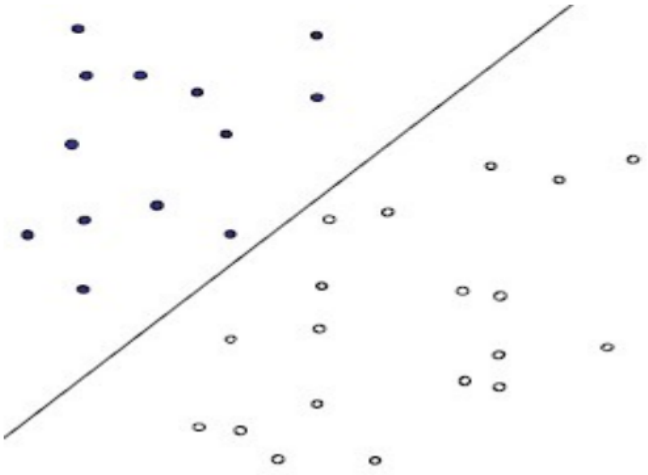
4.7.1 Mathematical Foundation

Linear classifier: Following figure shows a simple linear classification sample, in which a straight line is enough to classify two different types of nodes The linear classifier is defined as

follows.

(4.4)

where x, w, and b are vectors, and x is the input. The classifier f(x,w,b) has two possible values: -1 and +1. When x is located in a two-dimensional space, f(x) stands for a straight line. In a three dimensional space, f(x) stands for a plane surface. Generally, when x has a dimensionality n, f(x) is an n-1 dimensional hyperplane. From a linear classification problem, shown in Figure 3.1, however, there exist multiple classifiers (linear classifiers) to classify the given data nodes. The question is how to find out the ideal one. In SVM, the classifier which has the maximum margin is the one we are eager to find. Margin stands for the distance between the linear classifier and the nearest nodes, shown in following figure [26].



From Figure 3.2, the margin in (a) is obviously bigger than the margin in (b). In SVM, larger margin means better classification. The formula for calculating the margin (M) is defined as:

$$M = \frac{2}{\sqrt{w.w}} \tag{4.5}$$

4.7.2 Working Procedure of SVM

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. [27] A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane.

4.7.3 Different Kernels of SVM

The mathematical function used for the transformation is known as the kernel function.

The kernel function can be Linear,Polynomial,RBF,Sigmoid.

It represents a dot product of input data points mapped into the higher dimensional feature space by transformation [28]. Gamma is an adjustable parameter of certain kernel functions.

A linear kernel function is recommended when linear separation of the data is straightforward. In other cases, one of the other functions should be used. There need to experiment with the different functions to obtain the best model in each case, as they each use different algorithms and parameters.

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

4.8 Conclusion

We have applied the logistic regression and the decision boundary is linear that classify our dataset.In the case of support vector machine , linear kernel function make the hypothesis to separate the sentiment classification.

Chapter 5

Implementation

5.1 Introduction

In this chapter solution procedure of twitter sentiment analysis is done by using logistic regression and support vector machine will be discussed bases on theoretic discussion .The overall procedure is discussed in this chapter step by step and theinput output relation is also shown. At first the environment of implementation will be discussed.

5.2 Tools for Implementation

To implement this work , windows operating environment is used.Processor of 2.4 GHz core i3 and RAM of 4GB are used to produce result by this solution procedure describe bellow. Other tools are discussed below.

5.2.1 Python

Python is a high level, interpreted programming language, created by Guido van Rossum. The language is very popular for its code readability and compact line of codes. It uses white space inundation to delimit blocks. Python provides a large standard library which can be used for various applications for example natural language processing, machine learning, data analysis etc. It is favored for complex projects, because of its simplicity, diverse range of features and its dynamic nature.

5.2.2 Natural Language Processing (NLTK)

Natural Language toolkit (NLTK) is a library in python, which provides the base for text processing and classification. Operations such as tokenization, tagging, filtering, text manipulation can be performed with the use of NLTK. The NLTK library also embodies various trainable classifiers (example – Naïve Bayes Classifier). NLTK library is used for creating a bag-of words model, which is a type of unigram model for text. In this model, the number of occurrences of each word is counted. The data acquired can be used for training classifier models. The sentiment of the entire tweets is computed by assigning subjectivity score to each word using a sentiment lexicon.

5.2.3 SCIKIT-LEARN

The Scikit-learn project started as scikits.learn, a Google Summer Code project by David Cournapeau. It is a powerful library that provides many machine learning classification algorithms, efficient tools for data mining and data analysis. Below are various functions that can be performed using this library: [29]

- Classification: Identifying the category to which a particular object belongs.
- Regression: Predicting a continuous-valued attribute associated with an object.
- Clustering: Automatic grouping of similar objects into sets.
- Dimension Reduction: Reducing the number of random variables under consideration.
- Model selection: Comparing, validating and choosing parameters and models.
- Preprocessing: Feature extraction and normalization in order to transform input data for use with machine learning algorithm. In order to work with scikit-learn, we are required to install NumPy on the system.

5.2.4 NumPy

NumPy is the fundamental package for scientific computing with Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. It contains among other things:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities.

5.3 Setting Up Environment for Sentiment

Analysis Using Python The following components are required to be downloaded and installed properly.

- Download and install Python 2.6 or above in a desired location.
- Download and install NumPy.
- Download and install NLTK library.
- Download and install Scikit-learn library.

5.4 Dataset for Implementation

Twitter Sentiment analysis dataset is downloaded from kaggle . It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive) and they can be used to detect sentiment . It contains the following 6 fields:

target: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive) ids: The id of the tweet (2087) date: the date of the tweet (Sat May 16 23:58:44 UTC 2009) flag: The query (lyx). If there is no query, then this value is NO-QUERY. user: the user that tweeted (robotickilldozr) text: the text of the tweet (Lyx is cool) The official link regarding the dataset with resources about how it was generated is http://help.sentiment140.com/for-students/ . we use about 10,000 user information from this dataset.

5.5 Implementation Step

Implementation step in python are as follow:

5.5.1 Data preprocessing

Preprocessing includes the following:

- HTML decoding
- UTF-8 BOM (Byte Order Mark)
- removing numbers and special characters
- lower-case
- tokenizing and joining
- negation handling
- Removal of URLs. Frequently tweets contain web links to share some additional information. The content of the links is not analyzed, hence address of the link itself does not provide any useful information and its elimination can reduce the feature size, which is why URL is removed from the tweet.
- Removal of usernames. Another user can be mentioned by post creator in the tweet by using "@" symbol followed by username, i.e. @Superman. Due to this feature does not provide any relevant information it was also excluded from the tweet.
- Removal of hashtags. The hashtag is depicted using "# "symbol and used before a word that represents a topic name. Topics are not the task to be classified, hence they are omitted.
- Removal retweets and duplicates. The retweet is a tweet that is written by one user and then copied and posted by another user. Retweet contains "RT" abbreviation. Repeated tweets and retweets are removed in order to exclude putting extra weight on a specific tweet.

5.5.2 Data preparation

Before train any model, we first split the data. We chose to split the data into three chunks: train, development, test. Train set: The sample of data used for learning Test set: The sample of data used only to assess the performance of a final model. The ratio I decided to split my data

is 75/25, 75training set, and 25In this case, only 25enough to evaluate the model and rene the parameters. Another approach is splitting the data into only train and test set, and run k-fold cross validation on the training set, so that you can have an unbiased evaluation of a model. But considering the size of the data, I have decided to use the train set only to train a model, and evaluate on the dev set, so that I can quickly test dierent algorithms and run this process iteratively.

5.5.3 Feature generation

If we want to use text in machine learning algorithms, we'll have to convert them to a numerical representation. One of the methods is called bag-of-words approach. The bag of words model ignores grammar and order of words. Once we have a corpus (text data) then rst, a list of vocabulary is created based on the entire corpus. Then each document or data entry is represented as numerical vectors based on the vocabulary built from the corpora. With count vectorizer, we merely count the appearance of the words in each text. For example, let's say we have 3 documents in a corpus: "I love dogs", "I hate dogs and knitting", "Knitting is my hobby and my passion". If we build vocabulary from these three sentences and represent each document as count vectors, it will look like below pictures.

5.5.4 Basic Term for token analysis

Token Frequency:

Token is an individual occurrence of a linguistic unit in speech or writing. Token Frequency is the total number of words (tokens) in one or several texts .

Harmonic mean of frequency:

Harmonic Mean is a kind of average. To find the harmonic mean of a set of n numbers, add the reciprocals of the numbers in the set, divide the sum by n, then take the reciprocal of the result. The harmonic mean of a1, a2, a3, a4, . . ., an is given below.

$H = \frac{n}{\frac{1}{a_1}+\frac{1}{a_2}+\frac{1}{a_3}+...}$

Normalized cumulative frequency of all token: [30]

Let X is a random variable . X takes on numerical values between negative infinity and infinity. Now a question is "How likely is it that X will not exceed w?" where w can be Any real number. Once we ask this question, though, we are no longer looking for a single number

to describe the probability of an event, we are now looking for a (possibly) different number for EVERY single value of w between negative infinity and infinity.

This collection of numbers (each of which is the probability of a different event of the form $X <= w$) is called the cumulative probability function. It maps every real number w to a probability between 0 and 1. Once you find this function, you can just plug in the w that you want, and the function spits out the associated probability.

In textbooks and classrooms, the Cumulative Distribution Function is typically written as: $F(x) = P(X <= x)$

In my explanation, I used w in place of x because newcomers to the field are often confused when they see a X (i.e. the random variable) and x (i.e. any real number) in the same expression. F(w)=P(X¡=w) is equivalent to the usual textbook definition.

5.5.5 New Feature generation

In this section ,we will discuss the process of find a new feature. Feature is generated by finding the relation between token of different classes.Now we will try to find ,How different the tokens in two different classes(positive, negative). This time, the stop words will not help much, because the same high-frequency words (such as “the”, “to”) will equally frequent in both classes. If these stop words dominate both of the classes, I won’t be able to have a meaningful result. So, I decided to remove stop words, and also will limit the max features to 10,000 with countvectorizer.Now we will analysis the token. First, Top tokens in the dataset are generate . Following figure show the top 500 tokens.

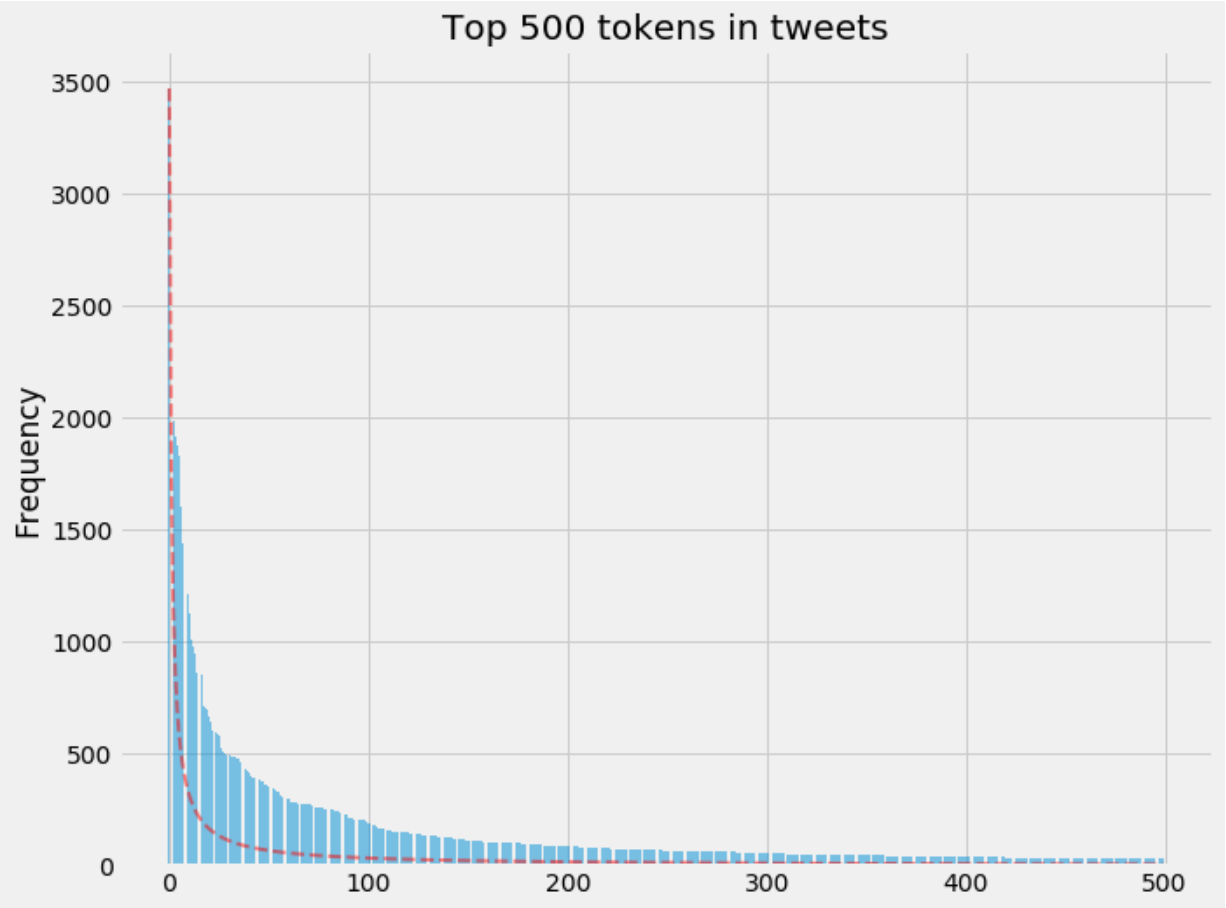


Figure 5.1: Diagram of Top tokens in tweet Dataset

Following figure show the top 50 tokens of positive class

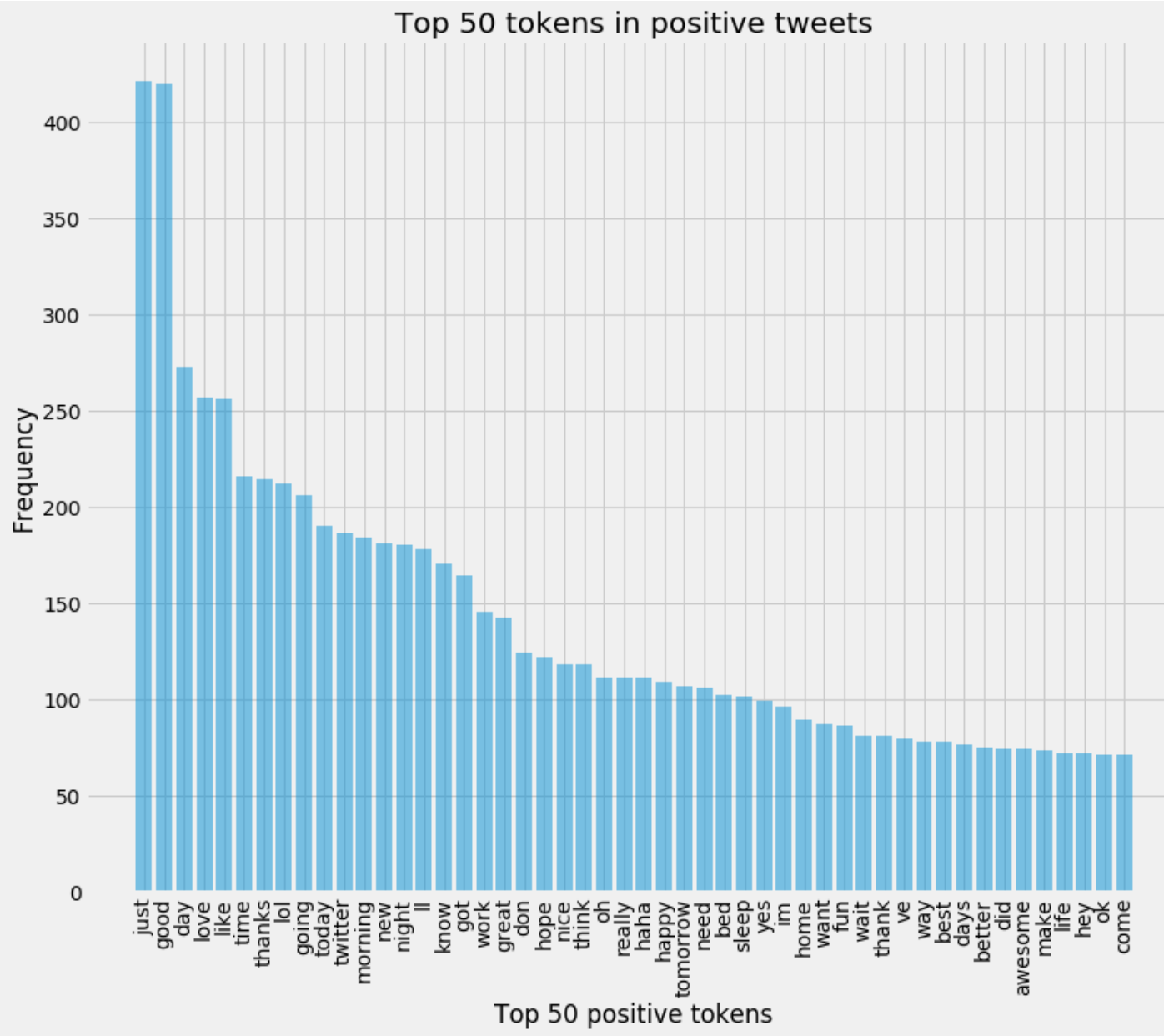


Figure 5.2: Diagram of Top tokens in tweet Dataset

Even though some of the top 50 tokens can provide some information about the negative tweets, some neutral words such as “just”, “day”, are one of the most frequent tokens. Even though these are the actual highfrequency words, but it is dicult to say that these words are all important words in negative tweets that characterises the negative class.Following figure show the top 50 tokens of negative class.

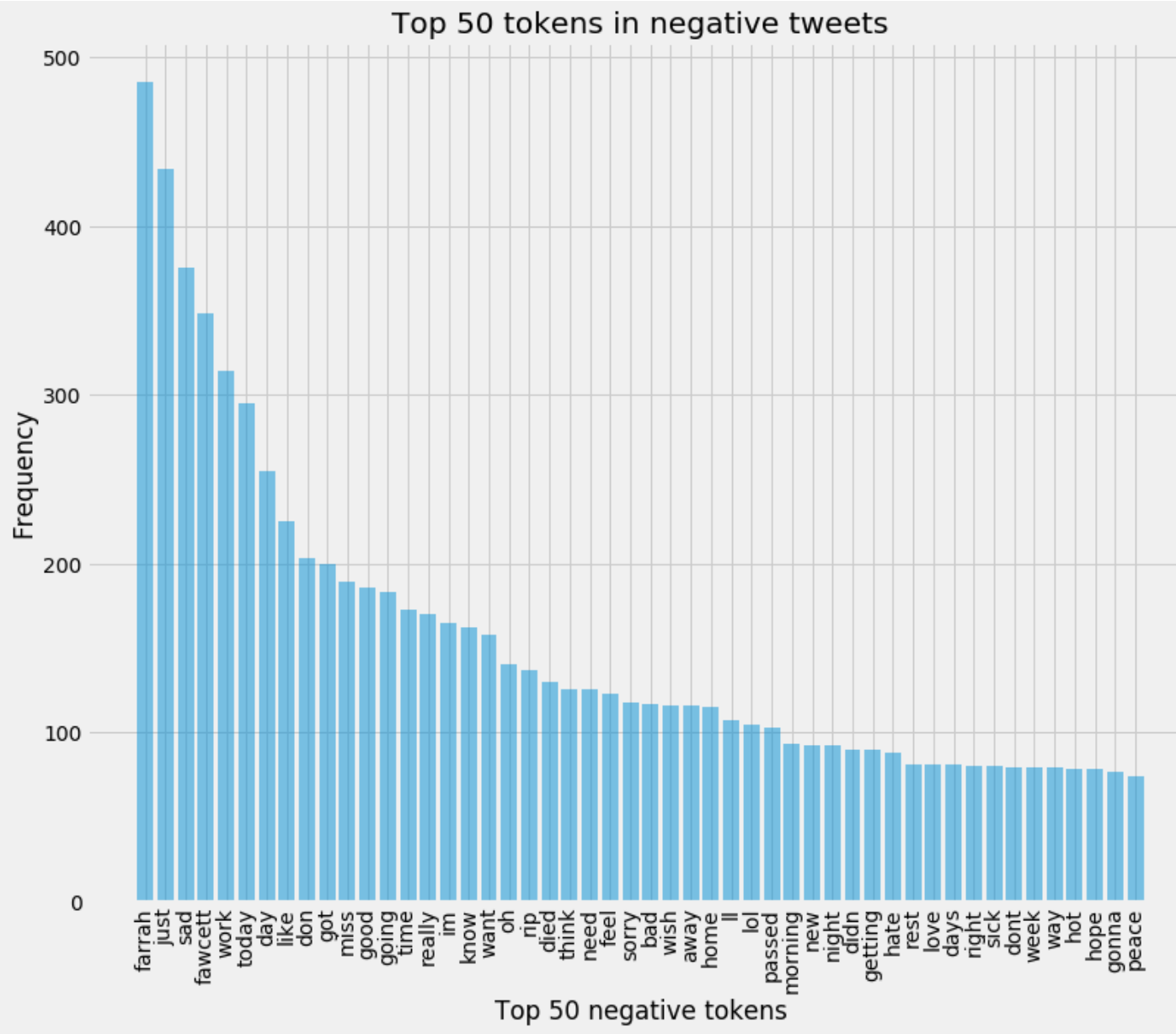


Figure 5.3: Diagram of Top tokens in tweet Dataset

Following figure show the relationship graph of frequency negative tokens vs positive tokens.

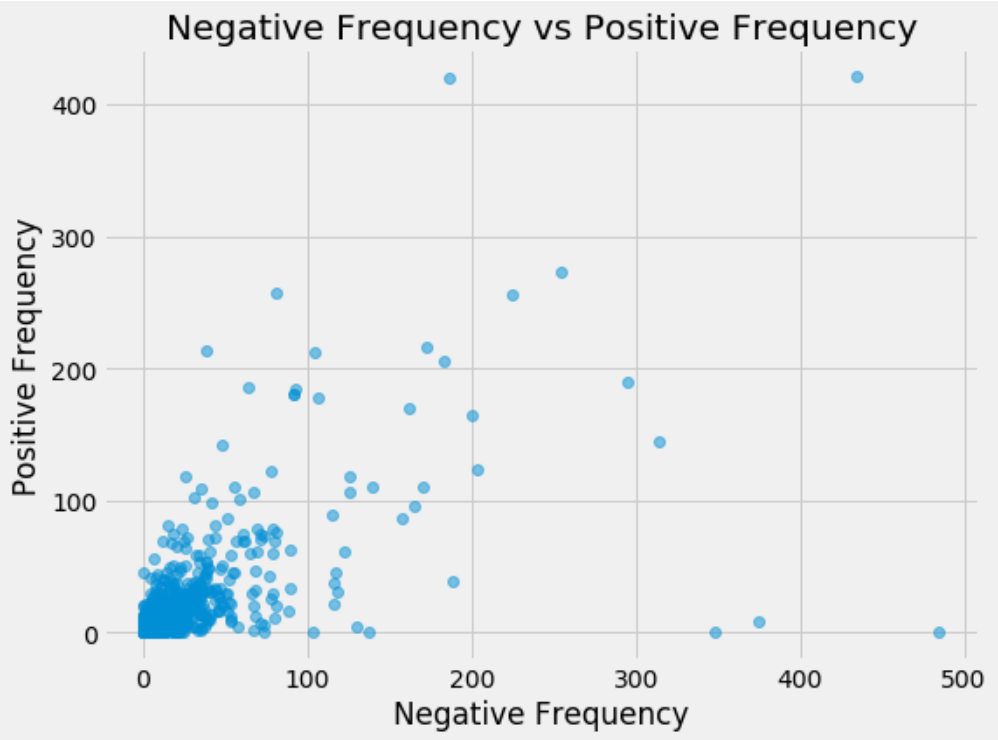


Figure 5.4: Diagram of Top tokens in tweet Dataset

Most of the words are below 10,000 on both X-axis and Y-axis, and we cannot see meaningful relations between negative and positive frequency.

Following figure show the relationship graph of harmonic mean of frequency of negative tokens vs positive tokens.

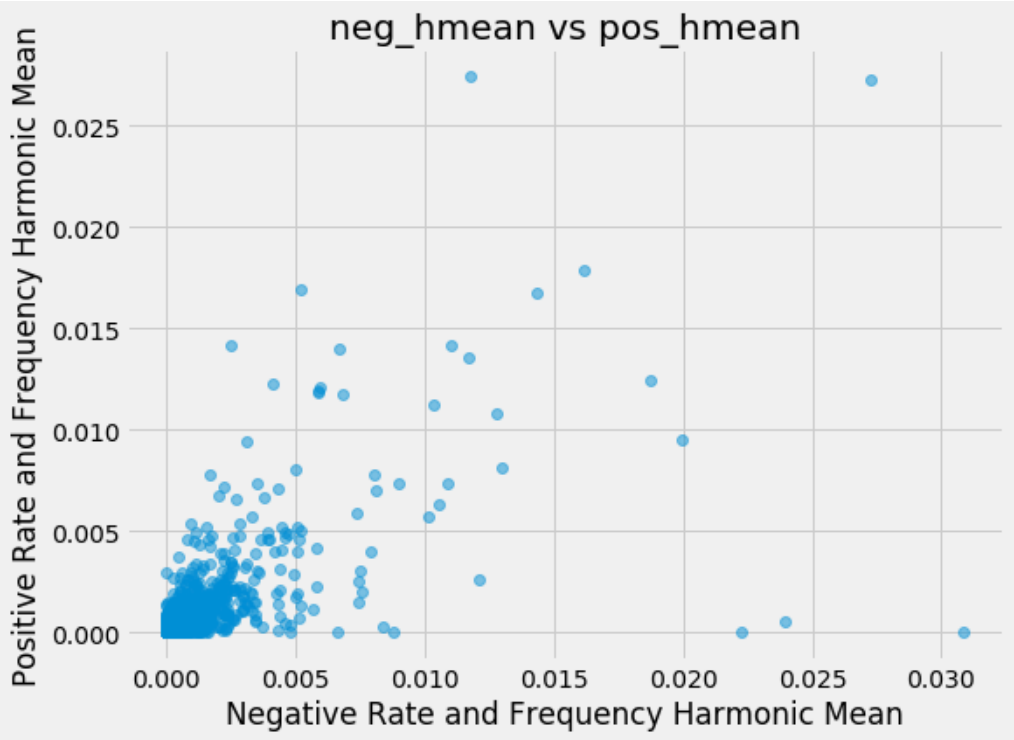


Figure 5.5: Diagram of Top tokens in tweet Dataset

Following figure show the relationship graph of normalized cdf of frequency of negative tokens vs positive tokens.

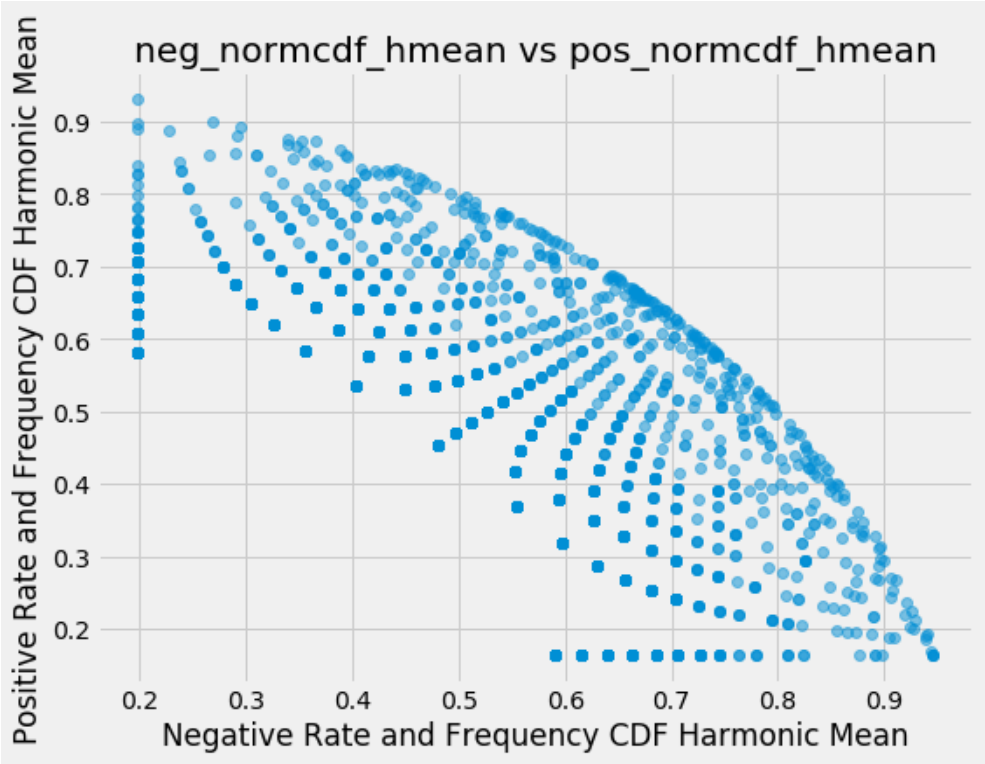


Figure 5.6: Diagram of Top tokens in tweet Dataset

Relation between Positive and Negetive class tokens From the previous analysis of token , we have seen that the diagram of normalized Cumulative Distribution Function of token can find a pattern likely to a linear negative slope .This means that if a data point is near to the upper left corner, it is more positive, and if it is closer to the bottom right corner, it is more negative.

Best feature: so we can use normalized Cumulative Distribution Function of frequency as a feature for sentiment analysis.

5.5.6 Combine Feature

Here all the feature for sentiment analysis.

Index	total	pos_rate	os_freq_p	pos_hmean	pos_rate_normcdf	s_freq_pct_norm	os_normcdf_hmea	neg_rate	neg_freq_pct	neg_hmean	neg_rate_n
glass	4	0.5	6.6613…	0.000133209	0.472144	0.468098	0.470112	0.5	6.45765e-05	0.000129136	0.527856
glasses	7	0.142…	3.3306…	6.65978e-05	0.177647	0.436475	0.252518	0.857143	0.00019373	0.000387372	0.822353
glasshouse	2	0	0	0	0.102727	0.405254	0.163906	1	6.45765e-05	0.000129145	0.897273
glasto	7	0	0	0	0.102727	0.405254	0.163906	1	0.000226018	0.000451934	0.897273
glastonbury	3	0	0	0	0.102727	0.405254	0.163906	1	9.68648e-05	0.000193711	0.897273
glau	2	1	6.6613…	0.000133218	0.870002	0.468098	0.608693	0	0	0	0.129998
globe	3	1	9.9920…	0.00019982	0.870002	0.499924	0.634975	0	0	0	0.129998
gloomy	2	0.5	3.3306…	6.66089e-05	0.472144	0.436475	0.453609	0.5	3.22883e-05	6.45724e-05	0.527856
glorious	4	0.5	6.6613…	0.000133209	0.472144	0.468098	0.470112	0.5	6.45765e-05	0.000129136	0.527856
gma	2	0.5	3.3306…	6.66089e-05	0.472144	0.436475	0.453609	0.5	3.22883e-05	6.45724e-05	0.527856
gmail	6	0.166…	3.3306…	6.66001e-05	0.19286	0.436475	0.267516	0.833333	0.000161441	0.00032282	0.80714
gnite	3	0.666…	6.6613…	0.000133213	0.628876	0.468098	0.536705	0.333333	3.22883e-05	6.45703e-05	0.371124
goal	3	0.666…	6.6613…	0.000133213	0.628876	0.468098	0.536705	0.333333	3.22883e-05	6.45703e-05	0.371124
god	80	0.375	0.0009…	0.00199309	0.356081	0.984459	0.522994	0.625	0.00161441	0.00322051	0.643919

Figure 5.7: Diagram of feature for sentiment analysis.

5.6 Training classifier model and Testing

The scikit-library provides various machine learning models whose implementation in code is very easy. For example one can easily create an instance of Support Vector Machine in one line : classifier=svm.SVC() In order to make use of machine learning models, one is required to remember to install NumPy properly and import from scikit-learn the desired model. After training the model we, use the same instance to test the model and save the results obtained.

The instance of logistic classifier is also create by: classifier=LogisticRegression()

5.7 Conclusion

In this chapter we preprocessed the dataset and then analysis the token relationship . we find the frequency,harmonic frequency and other data for those token.

Chapter 6

Results and Performance analysis

6.1 Introduction

This chapter introduces the results that were obtained after conducting the experiments using Logistic Regression and Support Vector Machine. The experiment is performed on the sentiment 140 datasets . The dataset is created from twitter.

6.2 Evaluation Measurement

The effectiveness of the classification algorithms is usually estimated based on such metrics as precision, recall, F score, and accuracy. Moreover, it is very important to take into account computational cost resources that algorithm needs for building the classifier and using it. Consider the metrics that were used for calculation of the precision, recall, F score, accuracy . Confusion matrix contains the estimated and actual distribution of labels. Each column corresponds to the actual label and each row corresponds to the estimated label of the sentence.

6.2.1 Confusion Matrix

TP is the number of true positives: the sentence that is actually positive and was estimated as positive, TN is the number of true negatives: the sentence that is actually negative and was estimated as negative, FP is the number of false positives: the sentence that is actually negative but estimated as positive, FNis the number of false negatives: the sentence that is actually positive but estimated as negative.

Table 6.1: Confusion matrix for a binary classifier

Actual / Estmated	Positive	Negative
Positive	TP	FP
Negative	FN	TN

6.2.2 precision

Precision can be estimated using following formula: $Precision = \frac{TP}{TP+FP}$ Precision shows how many positive answers that received from the classifier are correct. The greater precision the less number of false hits.

6.2.3 recall

However, precision does not show whether all the correct answers are returned by the classifier. In order to take into account the latter recall is used: $Recall = \frac{TP}{TP+FN}$ Recall shows the ability of the classifier to "guess" as many positive answers as possible out of the expected.

6.2.4 F1 Measure

The more precision and recall the better. However, simultaneous achievement of the high precision and recall is almost impossible in real life that is why the balance between two metrics has to be found. F_1 score is a harmonic mean of precision and recall:

$F_1 = \frac{2*precision*recall}{precision+recall}$

6.2.5 Accuracy

Accuracy (Acc) is the ratio of a total number of correctly classified samples (C) and the total number of samples (N). It varies between 0 (least accurate) and 1 (most accurate). If the accuracy is 1 that means the predictor is best. For calculating the performance of our system, we applied 5-fold cross-validation. For calculating the accuracy of our system, we use Equation 6.1.

$$Acc = \frac{C}{N} \tag{6.1}$$

Accuracy presents the proportion of the correct answers that are given by the classifier hence it can be estimated as: $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

6.3 Experimental Result

To access the classification performance ,evaluation criteria is a key factor.Table 6.2 demonstrates the accuracy of logistic classifier with unigram feature selection .

6.3.1 logistic classifier

Table 6.2: Accuracy of logistic classifier with unigram (c=10)

No. of frature	Traning Accuracy (%)	Testing Accuracy (%)
maximum feature=1000	82.71	75.23
maximum feature=5000	93.72	76.87
maximum feature=10000	96.41	76.95
maximum feature=20000	96.94	77.15
maximum feature=25000	96.94	77.15

Next we have worked with taking maximum feature=20000. Table 6.3 demonstrates the accuracy of logistic classifier with unigram feature selection varying the value of C .

Table 6.3: Accuracy of logistic classifier with unigram(maximun feature=20000)

Value of C	Traning Accuracy (%)	Testing Accuracy (%)
C=.01	77.33	73.10
C=.1	80.19	74.42
C=1	87.61	76.71
C=10	96.94	77.15
C=100	99.45	74.66

Table 6.4 demonstrates the accuracy of logistic classifier with bigram feature selection varying the value of C .

Table 6.4: Accuracy of logistic classifier with bigram(maximun feature=2000)

Value of C	Traning Accuracy (%)	Testing Accuracy (%)
C=.01	80.53	73.26
C=.1	82.81	74.74
C=1	90.80	76.67
C=10	99.37	70.03
C=100	99.81	75.87

So ,the maximum accuracy of logistic classifier is 77.15

6.3.2 SVM classifier

Now ,we use the svm classifier in our experiment .Table 6.5 demonstrates the accuracy of svm classifier varying the value of C .

Table 6.5: Accuracy of linear svm classifier

Value of C	Traning Accuracy (%)	Testing Accuracy (%)
C=.001	75.35	76.51
C=.01	83.53	82.93
C=.1	94.41	85.18
C=1	98.92	82.01
C=10	99.78	76.95
C=100	99.70	74.94

So ,the maximum accuracy of svm classifier is 85.18

Table 6.6 demonstrates the accuracy of svm classifier varying the value of C .

Table 6.6: Accuracy of modified Support Vector Machine classifier

Value of C	Traning Accuracy (%)	Testing Accuracy (%)
C=.001	81.26	74.74
C=.01	91.24	75.59
C=.1	96.62	74.3
C=1	98.94	71.37
C=10	99.62	71.29

6.4 Comparision

Table 6.7 demonstrates the comparision of previous result and our experimented results.

Table 6.7: omparision the result

Method	Previous Accuracy (%)	Experimented Method	Max-Accuracy (%)
SVM with unigram	82.2	logistic classifier unigram	77.15
SVM with bigram	78.8	logistic classifier bigram	76.67
SVM with unigram and bigram	81.6	svm–1 / svm-2	85.18 / 75.59

6.5 Conclusion

This chapter presents the results of conducted experiments using Logistic Regression and Support Vector Machine classifiers. It can be observed that Support Vector Machine gave quite good results then logistic regression.

Chapter 7

Conclusion and Future Works

7.1 Conclusion

The foremost purpose of the thesis work is develop a classifier that can more accurately classify the tweet in positive sentiment.For classifying ,firstly a new feature is determined by analysis the frequency of token and many mathematical evaluation of frequency is calculated to find the most relevant feature .

Then we use the new feature in our experiment and find a good result . The support vector machine classifier give a good performance then logistic classifier. we have proposed a system which uses supervised machine learning algorithm . Tweets can be classified based on the semantics they carry. This will reduce the overhead of traversing through various sites. Our system currently focuses on only text classification. In the future, the system can be trained to process emoticons in reviews. Different languages can be incorporated further into the system. The efficiency of the algorithm can be increased by using ensemble methods.

Machine learning approaches become an effective and popular tool to the area of sentiment analysis on tweets. Because of the unique characteristics of the tweet data, choosing machine learning classifiers and adjusting the parameters of algorithms are the essential tasks in the process of tweet sentiment analysis. On the other hand, since tweets are short and messy messages, the sentiment analysis methods that may work well in other data, such as movie reviews or pro duct reviews, are probably not effective for analyzing tweets. So identifying useful preprocessing methods plays an important role as well as the sentiment analysis algorithms. In our study, we carried out a systematic and thorough empirical study on the machine learning algorithms for tweet sentiment analysis. Through our experiments, we systematically compared

2 machine learning algorithms, and investigated the performance and effectiveness of these algorithms combined with some pre-processing methods, in terms of a variety of evaluation metrics. We found that the Support Vector Machine (SVM) was more effective and provided the best performance based on the experimental results. We specifically recommend SVM.

Secondly, we also consider the preprocessing method .Feature selection is very important for accurately classify the text sentiment.We have find feature by frequency analysis and apply it in classification algorithm.

7.2 Future Works

Future work will involve investigation of other approaches for preprocessing tweets because they have to be more thoroughly filtered to achieve the higher accuracy,precision, etc.There are several directions that can be performed: Tweets may contain a lot of spelling mistakes, hence, spelling corrector can be applied to exclude typos.Additionally, tweets contain huge amount of emoticons and expressions that convey laugh, such as lol, ha-ha-ha, jaja that have to be generalized and labeled whether emoticon/expression refers to a positive or negative meaning, the ones that are ambiguous (e.g. emoticon with stuck-out tongue " :-P ") have to be removed from the training dataset. Another experiment that may be carried out is the replacement of the abbreviations with their full meaning. It obviously will increase the size of the training corpus but may add more sense to the tweet.Moreover, it would be interesting to add neutral class and check the performance of the classifier. However, in this case, the training and testing datasets have to include neutral samples to feed the model and evaluate it.

REFFERENCES

[1] A. B. Archive, "Introduction to sentiment analysis."

[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, pp. P5(1),1–167, 2012.

[3] D. Zimbra, M. Ghiassi, and S. Lee, "Brand-related twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks," in *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. IEEE, 2016, pp. 1930–1938.

[4] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, "Opinion mining and sentiment analysis on a twitter data stream," in *Advances in ICT for emerging regions (ICTer), 2012 International Conference on*. IEEE, 2012, pp. 182–188.

[5] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, no. 12, 2009.

[6] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[7] S. Gopinath, "Types of sentiment analysis," 2014.

[8] E. H. Hovy, "What are sentiment, affect, and emotion? applying the methodology of michael zock to sentiment analysis," in *Language production, cognition, and the Lexicon*. Springer, 2015, pp. 13–24.

[9] A. Pawar, M. Jawale, and D. Kyatanavar, "Fundamentals of sentiment analysis: Concepts and methodology," in *Sentiment Analysis and Ontology Engineering*. Springer, 2016, pp. 25–48.

[10] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[11] R. Feldman, J. Sanger *et al.*, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[12] P. Haseena Rahmath and T. Ahmad, "Sentiment analysis techniques-a comparative study," 2014.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[14] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003, pp. 519–528.

[15] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics*, vol. 3, no. 2, pp. 143–157, 2009.

[16] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 12, 2008.

[17] O. Kolchyna, T. T. Souza, P. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," *arXiv preprint arXiv:1507.00955*, 2015.

[18] M. Annett and G. Kondrak, "A comparison of sentiment analysis techniques: Polarizing movie blogs," in *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2008, pp. 25–35.

[19] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[20] A. Mudinas, D. Zhang, and M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis," in *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*. ACM, 2012, p. 5.

[21] S. Trinh, L. Nguyen, and M. Vo, "Combining lexicon-based and learning-based methods for sentiment analysis for product reviews in vietnamese language," in *International Conference on Computer and Information Science*. Springer, 2017, pp. 57–75.

[22] T. R. Ricky, "another-twitter-sentiment-analysis," january,2013.

[23] J. D'Souza, "introduction-to-bag-of-words," Apr 3.

[24] P. Chandrayan, "Machine learning part 3 : Logistic regression," *Towardsdatascience.com*.

[25] S. Narkhede, "Understanding logistic regression," *Towardsdatascience. com*.

[26] C. Fernandez-Lozano, E. Fernández-Blanco, K. Dave, N. Pedreira, M. Gestal, J. Dorado, and C. R. Munteanu, "Improving enzyme regulatory protein classification by means of svm-rfe feature selection," *Molecular Biosystems*, vol. 10, no. 5, pp. 1063–1071, 2014.

[27] S. R. Gunn *et al.*, "Support vector machines for classification and regression," *ISIS technical report*, vol. 14, no. 1, pp. 5–16, 1998.

[28] B. Schölkopf, S. Mika, C. J. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[29] S. Raschka, *Python machine learning*. Packt Publishing Ltd, 2015.

[30] M. Lamar, "What is a cumulative distribution function?" 2018.