SENTIMENT ANALYSIS AND OPINION MINING VIA MICROBLOGGING IN SOCIAL MEDIA LIKE: TWITTER

MUSTAFA SALMAN ABD AL-BNDI

JANUARY 2015

SENTIMENT ANALYSIS AND OPINION MINING VIA MICROBLOGGING IN SOCIAL MEDIA LIKE: TWITTER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES OF
ÇANKAYA UNIVERSITY

BY
MUSTAFA SALMAN ABD AL-BNDI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF
MATHEMATICS AND COMPUTER SCIENCE/ INFORMATION
TECHNOLOGY PROGRAM

JANUARY 2015

Title of the Thesis: **Sentiment Analysis and Opinion Mining via Microblogging in Social Media Like: Twitter.**

Submitted by **Mustafa Salman Abd AL-BNDI**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.

Prof. Dr. Taner ALTUNOK
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Billur K

Prof. Dr. Billur KAYMAKÇALAN
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assist.Prof. Dr. Abdül Kadir GÖRÜR
Supervisor

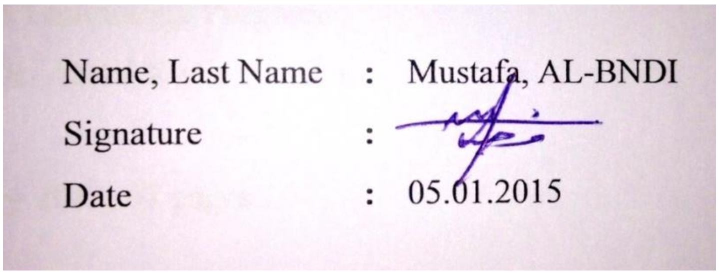
Examination Date: 05.01.2015

Examining Committee Members

Assoc. Prof. Dr. Ersin ELBAŞI	(İpek Univ.)	
Assist. Prof. Dr. Abdül Kadir GÖRÜR	(Çankaya Univ.)	
Assist. Prof. Dr. Tolga PUSATLI	(Çankaya Univ.)	

STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.



Name, Last Name : Mustafa, AL-BNDI

Signature :

Date : 05.01.2015

ABSTRACT

SENTIMENT ANALYSIS AND OPINION MINING VIA MICROBLOGGING IN SOCIAL MEDIA LIKE: TWITTER

AL-BNDI, Mustafa
M.Sc., Department of Mathematics and Computer Science/
Information Technology Program
Assist. Prof. Dr. Abdül Kadir GÖRÜR

January 2015, 67 pages

This research is a study of microblogging on social websites such as *Twitter* and shows the techniques of emotion detection and sentiment analysis for the same. This research has three objectives. The first objective is a discussion about how to extract and classify emotions in tweets using the unigram feature extractor with word presence or word frequency as a factor of extraction. High accuracy of classification is obtained when considering the word presence as a factor of extraction. Moreover, one can obtain high accuracy also by using word frequency as a factor of extraction when supplying the test data on training corpora of tweets in the case of multi-domain tweets. The second objective is the extraction and classification of the emotions of tweets using n-gram (1<n<4) feature extractors. We illustrate how to obtain high accuracy of classification through an increase in the number of instances of tweets for training corpora, and we prove that through supplying a test dataset on increasingly growing groups which are collected sequentially from training corpora with equal distribution of positive and negative tweets. A sentiment classifier has been used with models such as the Multi-nominal Naïve-Bayes model and the Sequential Minimal Optimisation, which is a type of Support Vector Machine model. Finally, we determine which one of two selected machine learning models is more suitable for classifying the sentiment of tweets in order to determine whether a tweet has a positive or negative sentiment.

Keywords: Sentiment Classification, Opinion Mining, MNNB, SMO, Word Presence, Word Frequency, Stemming, Multi-domain tweets, N-gram Feature Extractor.

ÖZ

TWİTTER GİBİ SOSYAL MEDYA ORTAMLARINDA MİKRO BLOGLAMA YOLUYLA DUYGU ANALİZİ VE FİKİR MADENCİLİĞİ

AL-BNDI, Mustafa
Yüksek Lisans, Matematik-Bilgisayar Anabilim Dalı/
Bilgi Teknolojisi Bölümü
Doçent Dr. Abdülkadir GÖRÜR

Ocak 2015, 67 sayfa

Bu çalışmada sosyal medyadaki *Twitter* gibi web sitelerinde bulunan mikro bloglama fonksiyonu araştırılmakta ve bu sitelerdeki duygu tarama ve duygu analizi teknikleri gösterilmektedir. Bu araştırmanın üç tane amacı vardır. Birinci amaç tweetlerdeki duyguların kelime bulma veya kelime sıklığı özelliklerini bir çıkarım faktörü olarak kullanan unigram özellik çıkarıcı uygulamasını kullanılarak nasıl çıkarılacağı ve sınıflandırılacağı konusunu irdelemektir. Kelime bulma özelliği bir çıkarım faktörü olarak dikkate alındığında yüksek bir doğruluk oranı elde edilir. Ayrıca, çok alanlı tweetlerde tweetlerin eğitim korporasına test verileri verilirken kelime sıklığını bir çıkarım faktörü olarak kullanarak da yüksek doğruluk oranı elde edilebilir. İkinci amaç tweetlerdeki duyguların n-gram (1<n<4) özellik çıkarıcıları kullanılarak çıkarılması ve sınıflandırılmasıdır. Eğitim korporasında tweet örneklerinin sayısındaki bir artış yoluyla sınıflandırmada nasıl yüksek bir doğruluk oranı elde edileceğini gösteriyoruz ve bunu eşit sayıda pozitif ve negatif tweet dağılımıyla eğitim korporasından sıralı olarak toplanan artan biçimde büyüyen gruplara bir test dataseti vermek suretiyle kanıtlıyoruz. Bir Destek Vektörü Makine Modeli türü olan Sıralı Minimal Optimizasyon ve Multi-nominal Naïve-Bayes modeli gibi modeller ile birlikte bir duygu sınıflandırıcısı kullanılmıştır. Son olarak, bir tweetin pozitif duyguya mı yoksa negatif duyguya mı sahip olduğunu belirlemek amacıyla tweetlerin duygularını sınıflandırmak seçilen iki makine öğrenme modelinden hangisinin daha uygun olduğunu belirliyoruz.

Anahtar Kelimeler: Duygu Sınıflandırması, Fikir Madenciliği, MNNB, SMO, Kelime Bulma, Kelime Sıklığı, Kök Bulma, Çok Alanlı Tweetler, N-gram Özellik Çıkarıcı.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Assist. Prof. Dr. Abdül Kadir GÖRÜR for his supervision, special guidance, suggestions, and encouragement through the development of this thesis. Also I cannot forget the best Instructor and advisor the Assist. Prof. Dr. Tolga PUSATLI, for his special guidance, suggestions, and encouragement through all Master study.

It is a pleasure to express my special thanks to my family for their valuable support, also to my beloved country Iraq.

TABLE OF CONTENTS

CHAPTERS:

LIST OF FIGURES

FIGURES

LIST OF TABLES

TABLES

LIST OF ABBREVIATIONS

MNNB	Multi-Nominal Naïve-Bayes
SMO	Sequential Minimal Optimisation
SVM	Support Vector Machine
USP	Unique Selling Proportion
API	Application Programming Interface
P.P	Pre-processing
T.N	Target Name
URL	Uniform Resource Locator
H.T.T	Hash Topic Tag
Punc.	Punctuation
Spec.	Special
Arith.	Arithmetic
S.L	Stop List
L.C	Lower Case
Stm.	Stemming
Pres.	word Presence
Freq.	word Frequency
F.R	Feature Reduction
Acc.	Accuracy
M.L.C	Machine Learning Algorithm

CHAPTER 1

INTRODUCTION

1.1 Overview

In the modern world microblogging has become popular. The same is also seen among the people wanting to share their opinions. Opinion mining and sentiment analysis can be said to have brought in a large amount of interest in present day studies. This is due to the difficulties in the study of trending analysis. Therefore, the early work occurred at the document level and by applying different methods used for classifying a document's polarity. This can be done for positive, negative or neutral emotions at any scale.

One procedure to yield knowledge is opinion mining. The same can be said for the opinions of people sharing on social websites, blogs, groups and comment boxes. Opinion mining uses text mining and natural language processing procedures so that a computer can take in the expression of emotions. Moreover, it is helpful in bringing out the sentiment and emotional expressions from unstructured text and providing the best method to classify a given sentiment analysis [1].

Past research in sentiment analysis includes the work done by **Pang** [2], who has assessed the workings of the various classifiers on movie reviews. This piece of research has been the baseline for further research. The techniques which have been discussed stretch across various specialised domains and **Pang** [2] has made use of the star ratings for the polarity signals when there is training data.

He shows comparable results for tweets with distant supervision. Moreover, to analyse the workings of a classifier, supervised learning is crucial. However, most of the time, a tweet and its sentiment cannot be recognised accurately. In addition, **Pang's** [2] research has made use of the litmus test such that if a person tweets and the same is seen on a front page newspaper headline, then it could be said to be in the neutral class.

Here, an example of the advantages of harnessing sentiment analysis to serve the needs of an organisation includes **IBM Cognos Consumer Insight Company** [3]. It represents the analytics platform, which is specially designed for social media. This platform uses advanced software to extract sentiments and trends which are mentioned in conversations among consumers. Examples include some products that span different social media. Therefore, one can use this information to guide a company to be more precise and to enhance its products so as to increase consumer conviction.

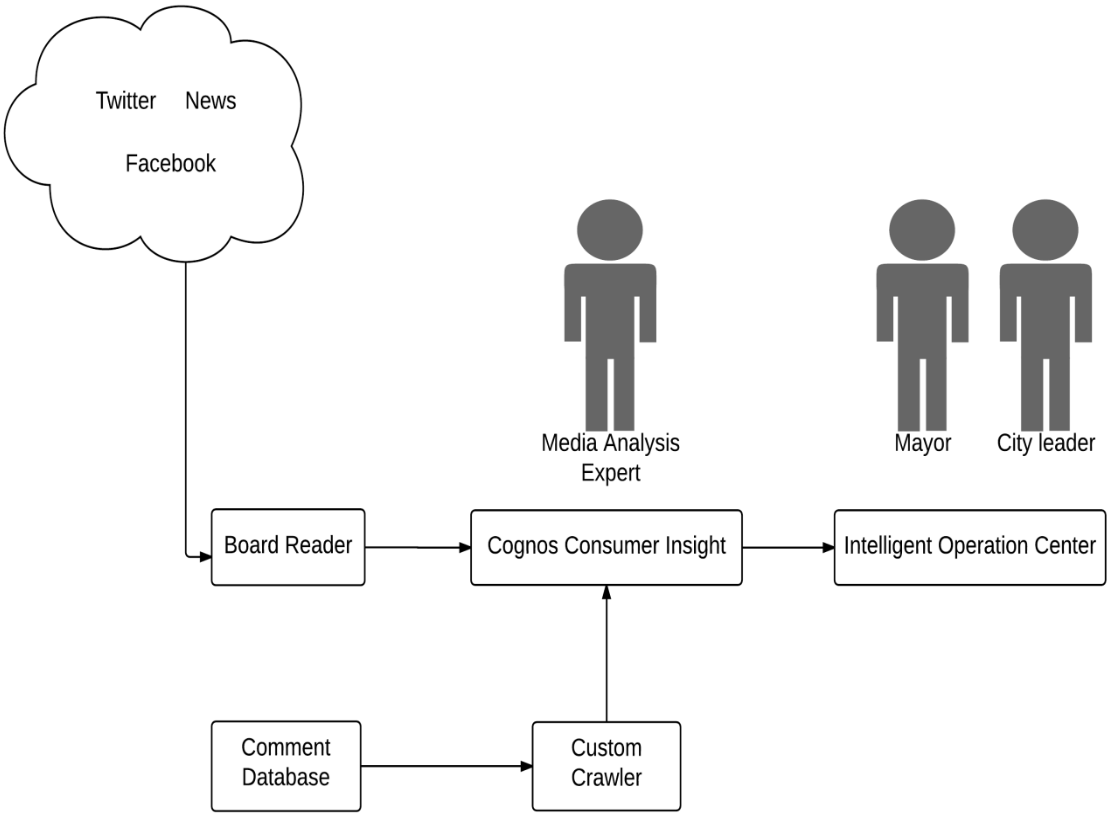


Figure 1 IBM Cognos consumer Insight Company

As can be seen in Figure 1 for the IBM Cognos consumer insight company [3], the process of sentiment starts with a social network site and then the board reader picks up the comments as the first stage. Media expert analysis is responsible for Cognos Consumer Insight as a second stage and finally the intelligence operation centre, which processes the post and puts out the result as either positive or negative under company consideration. The comment database is also crucial in this process because it forms a reference or guide to the new posts. Additionally, one would need a custom crawler which uses crawling or spider software to update their contents of comments which is then indexed to ensure quick processing by saving the visited comments and their authors for later processing.

1.2 Features of Tweets

As already known, the maximum length of a Twitter message is 140 characters and therefore many attributes can be seen in Twitter messages. These attributes include **USP** (Unique Selling Proportion) [4]. Furthermore, not all attributes will be handled as the features of tweets since most will not be useful. This also depends on research objectives that need to be accessed and how harnessing the model to serve these objectives in order to reach the threshold by obtaining the best accuracy of sentiment classification of the tweets which have been fetched from Twitter. The previous sentiment classification which most researchers stress is in the longer bodies of work on classification such as the case of movie reviews.

Data availability is also crucial in this context. The Twitter API has made it easy to collect millions of tweets for the purpose of training. Moreover, it has been seen in the previous research that the tests have studied thousands of training items [5].

The language model Twitter users use is brought in from the messages from various media. Messages may also come from smart cell phones. Another concern is the prevalence of misspellings and slang, which is higher on Twitter and Facebook in comparison to other domains [5].

Twitter users are used to posting short messages about a variety of topics. This is not the case among other sites which have been customised on a specific topic [6]. Therefore, a large percentage of past research has focused on specific data domains, such as movie reviews [7].

1.3 Sentiment Analysis or Mining of Regular Opinions

In this research, the main aim is the mining of online opinions. This is done by collecting various tweets, blogs, reviews, etc. from known social sites such as *Twitter*. Thus the interest of the research is to mine the features and aspects of the collected entities or tweets. This may include products or topics such as news, finance, movies, employment and sport for which people have expressed their opinions on the social website [6].

These opinions can be categorised under positive or negative segments. For summaries about an opinion, one can use the quantitative aspect. This is because about 50 percent of the time people opine that a product or service is good or bad according to **Jindal N. et al.** [7].

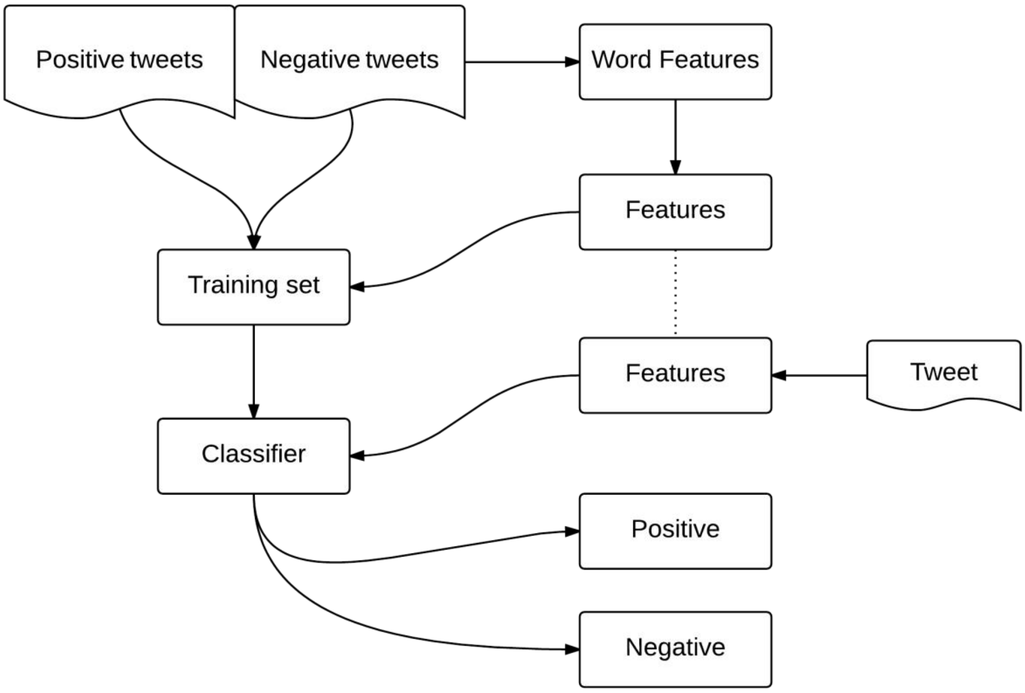


Figure 2 High level tweets classification

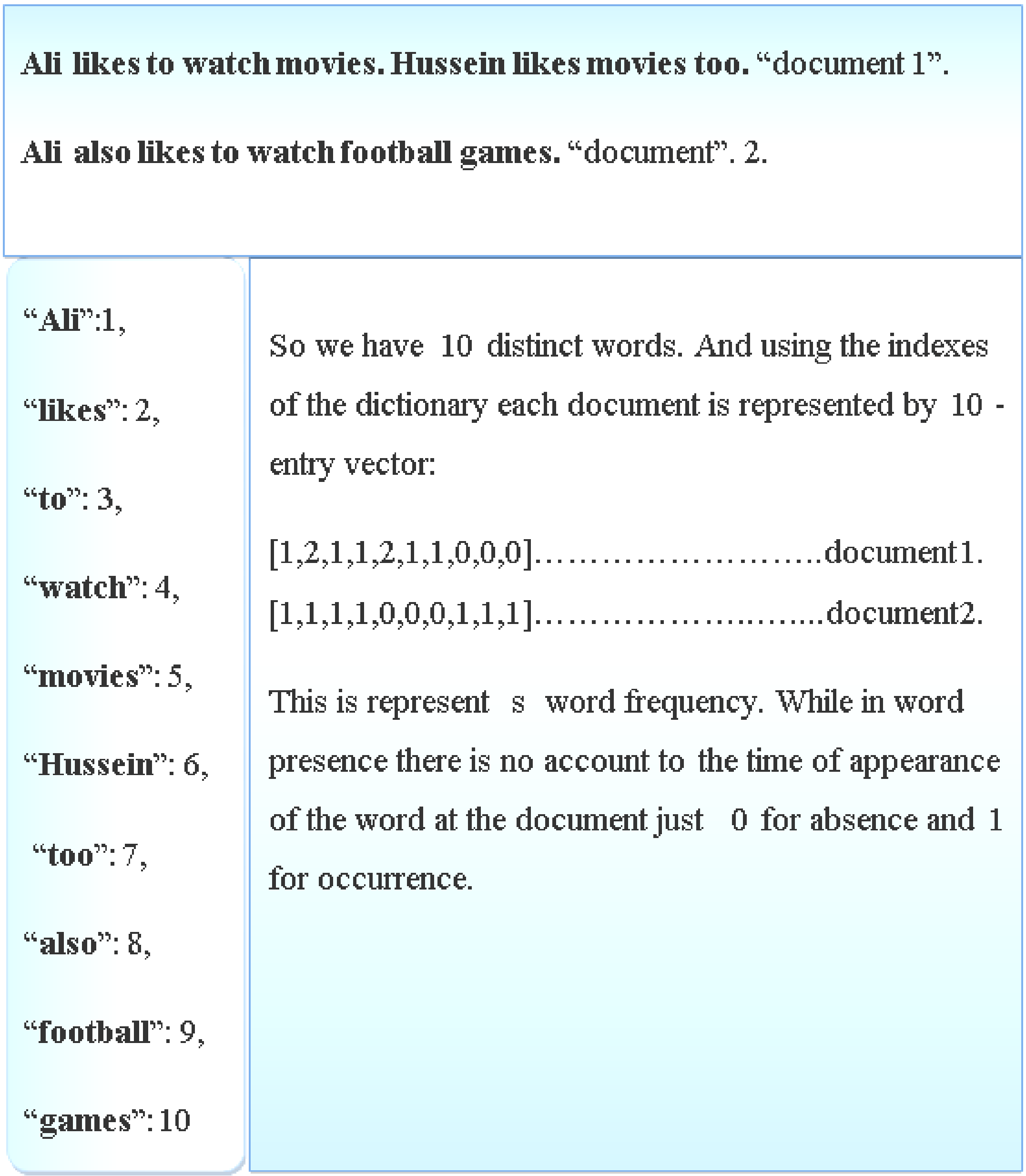
In the process, it is very important to classify positive and negative tweets, as shown in Figure 2 [8]. It is true that expressions can be picked up from the unstructured text. The next stage would be to find the best method to segregate the factors of sentiment analysis.

1.4 Word Presence vs. Word Frequency

Before beginning to explain our objectives, we need to know the meaning of presence and frequency of the words in a specified document. Therefore, we begin with word presence, which refers to a binary indicator for word occurrence and merely indicates whether or not a word occurs without capturing the number of times the word occurs. Moreover, the order of the words is lost. The same occurs with regard to word frequency except that the number of occurrences of words in a document is captured. Furthermore, these two approaches are widely used in text classification, such as the Multivariate Bernoulli document model and the Multinominal document model [9].

In the case of word presence, the document is presented as an event with word occurrence or absence being presented as an attribute of that document. On the other hand, word frequency is the occurrence of the individual word presented as an event with the document as a group of word events [9].

The following example clarifies the difference between word frequency and word presence in a document (Figure 3).



Ali likes to watch movies. Hussein likes movies too. “document 1”.

Ali also likes to watch football games. “document”. 2.

“**Ali**”:1,

“**likes**”: 2,

“**to**”: 3,

“**watch**”: 4,

“**movies**”: 5,

“**Hussein**”: 6,

“**too**”: 7,

“**also**”: 8,

“**football**”: 9,

“**games**”: 10

So we have 10 distinct words. And using the indexes of the dictionary each document is represented by 10 - entry vector:

[1,2,1,1,2,1,1,0,0,0]……………………..document1.
[1,1,1,1,0,0,0,1,1,1]…………………..…..document2.

This is represent s word frequency. While in word presence there is no account to the time of appearance of the word at the document just 0 for absence and 1 for occurrence.

Figure 3 Example of word frequency vs. word presence

1.5 Thesis Objectives

In addition to showing some considered techniques which are followed for extracting emotions and sentiments of tweets, this research also proposes three objectives. The first objective is to discuss how to extract and classify the emotions of tweets using the unigram feature extractor with word presence or word frequency as a factor of extraction. High accuracy of classification is obtained when considering word presence as a factor of extraction. Moreover, one can obtain high accuracy by using word frequency as a factor of extraction when supplying the test data on the training corpora of multi-domain tweets. While **B. Pang et al.** [10] was using the unigram feature extractor with word presence as a factor of extraction, high accuracy was obtained against using word frequency in the case of single-domain collected tweets, such as a movie domain.

The second objective is to extract and classify emotions of tweets using n-gram (1<n<4) feature extractors and to illustrate how to obtain a high accuracy of classification through an increase in the number of instances of tweets for training corpora, and to prove that through supplying the test dataset on increasingly growing groups which are collected sequentially from the training corpora with equal distribution between positive and negative tweets. The sentiment classifiers have been used with models such as MNNB (Multi-nominal Naïve-Bayes) for text categorisation and SMO (Sequential Minimal Optimization) which solves the quadratic programming problem detected through the training of SVM (Support Vector Machine).

Finally, we determine which one of the two selected machine learning models is more suitable for classifying the sentiment of tweets.

Here the time and complexity which has been used for classifying the sentimentality of the collected tweets will be reduced by the given approach. As well as the previous literature related to sentiment analysis, blogging and microblogging functions will be studied thoroughly.

1.6 Limitations

This research has two limitations, the first of which is the fact that neutral tweets have not been included in any training or testing data. The classification which is carried out here is for positive or negative tweets. Moreover, the stop word list which is included to be removed from the training corpora and test datasets, is a sample from the main default stop word list. This represents only the Google stop words which were inspired by Google a decade ago. Furthermore, a number of extra cases are added which do not affect the meaning of the sentences, including negation stop words. This is the second limitation of this research.

1.7 Thesis Structure

The literature review chapter examines the various methods in which the Twitter Corpora have been collected. Along with this, a linguistic analysis will be done on the corpus and the sentiment classifier will be analysed with some major experimental evaluations. The third chapter will type the paper methodology, training corpora, test dataset, pre-processing and feature reduction operations, machine learning types, the MNNB algorithm model and the SMO model. The results obtained from the two approaches will be discussed in the fourth chapter. The chapter of conclusions will list the major conclusions of the research and the challenges that have been made by applying our approach. Lastly, an Appendix of Tables (Chapter 6) illustrates the detailed results to be as guidance for researchers and students.

CHAPTER 2

BACKGROUND and REVIEW

2.1 Sentiment Analysis

Sentiment analysis, also known as **opinion mining,** is the use of natural language processing along with text analysis and computational linguistics. These are used together to identify the subjective information in source materials and to extract information from social websites, such as Twitter and Facebook.

2.2 Sentiment Analysis and Web 2.0

It is true that the fame of social media such as *Twitter*, *Facebook*, various blogs, etc. have given a lot of importance to sentiment analysis. Thus the proliferation of reviews, recommendations, ratings, related types of online expression and online opinions may be taken as the virtual currency for businesses. Moreover, sentiment may show the attitudes of customers towards products, thus giving companies the ability to identify new opportunities and determine the value of their actual reputations in the eyes of customers.

Many companies are attempting to automate the process of filtering. This will leave them with genuine blogs and filter out the noise. Thus, even at a commercial level, it becomes important to understand conversations and look at the relevant content and absorb its functioning properly [11].

For this reason, the field of sentiment analysis is crucial. Web 2.0 [12] is fully about the democratisation of publishing. Thus the web is now progressing towards being based on democratized data mining for published content [12]. With this aim in mind, one can go through many published studies. Many a research team at universities have been stressing the understanding of the ever-changing phenomenon of sentiment [13] so that sentiment analysis can help us to know why some electronic communities, such as *MySpace*, fade away while others maintain continuously high ratings and growth, such as *Facebook* and *Twitter*.

The Cyber Emotions project has analysed the role of negative emotions, which was motivated by social networks discussions. Here, the problem is that most sentiment analysis algorithms have been using simple terms for the expression of sentiment for a product [14].

There are, however, some crucial cultural factors and linguistic changes present among the different contexts, which presents difficulties when turning the written text to the related positive or negative sentiment [11]. Human beings disagree on the sentiment of text, which then becomes a greater task for computers to carry out sentiments classification. If a string of text is shorter, then it becomes more difficult to search for and assign a sentiment to it.

2.3 Previous Related Works

Some of the crucial research work present in the given subject is that of **Turney** [15] and **B. Pang et al**. [10], who used many methods to judge the polarity of many reviews, including product and movie reviews. The same has been done at the document level. Furthermore, there is a study about polarity, namely a multi-way scale that has been explained by **Pang** [2] and **Snyder** [16], who elaborated the basic task in classifying a movie review. These were further classified as positive or negative when they were discussing star ratings which could be seen at the 3- or 4-star scale.

At the same time, **Snyder** [16] had conducted an in-depth analysis which included restaurant reviews. For these reviews, he predicted ratings which included many aspects of the given restaurant. These included reviews of the food and atmosphere in relation to the five-star scale.

The same is also a statistical classification method in which the neutral class was not taken into account. Here the assumption in the research was that neutral texts will be present in the boundary of the binary classifier [17]. Researchers have not conducted much research on the polarity problem other than the three categories which have already been identified.

Additionally, another method used by many researchers is the scaling system in which words are categorised into negative, neutral or positive sentiments. They are measured on a scale of –10 to +10 (which ranges from the most negative to the most positive).

In case a piece of unstructured text is analysed with the use of natural language processing, it can be seen that the subsequent concepts are analysed for an understanding of the given words and their relation to the concept. The concept is then given a score for the sentiment to which the words relate and an associated score is given to this word. The texts can also be given a positive and negative sentiment, which is the strength score in case one has to show the sentiment of the text in comparison to the overall polarity along with the strength of the text [18].

The same can be said when classifying a given text which may be a sentence in one of two classes: either objective or subjective. This is an issue when the polarity classification is to be carried out. It can be said that the subjectivity of words will depend on the context .Thus an objective document may also have subjective sentences (such as news articles which discuss people’s opinions) [2]. In previous studies, the results rely on the definition of subjectivity, which is utilised when one is annotating texts. At the same time, **Pang** [2] discusses removal of objective sentences from a document. Thus, in an opinion mining context in order to determine the feature that is opinionated, the grammatical relationship used in the set of words is crucial and needs to be studied. Grammatical dependency relations analysis is conducted by the deep parsing of the text.

In the automated method, the Open Source software tools will make use of statistics, machine learning and natural language processing techniques for the sentiment analysis. The same is best suited for large collections of texts, which include web pages, online news, Internet discussion groups, web blogs, online reviews and social media [19].

It can be said that the structure of sentiments along with the topics can be very complex. There is also the problem of sentiment analysis, which could be non-monotonic in relation to the sentence extension along with stop-word substitution.

When we look at the field of *sentiment analysis and opinion mining*, the same can be related to social media. There is a large amount of material on the Internet, including reviews, blogs, tweets and forum discussions, which can be analysed for positive or negative sentiments. In previous research, there was a *Feature-Based Opinion Mining* [20] model proposed. This model is now known as *Aspect-Based Opinion Mining* [21]. Here it is clear in both the text that there is a degree of positive and negative emotion. The differentiation of emotion is very clear and thus the sentiment analysis in this case becomes easier.

Furthermore, some work has been conducted in the past on *Fake review and opinion spam detection*. Fake reviews could also be called bogus reviews or fraudulent reviews. These are difficult to detect and more development is required for accuracy in detection [22].

2.4 The Main Mining Tasks

- Mining entities and their features (or aspects) that have been commented on or evaluated by people.
- Determining whether the comment/opinion on each entity feature (or aspect) is positive, negative or neutral (aspect-based sentiment classification) [23].

2.5 Opinion Mining (OM)

Here, the focus is on opinion mining research which discusses people’s positive and negative sentiments. This is done to classify the sentiment for further research. Opinion mining, or sentiment mining, is a crucial area of research. The main aim is to design automatic systems which can determine human opinions which are prevalent from text written in natural languages. Thus, it can be said that textual information has two parts for classification which can be listed as either facts or opinions. **Facts** may be defined as objective phrases which list events and properties [7]. In contrast, **Opinions** are said to be subjective expressions. Therefore, it can also be concluded that the person may express positive or negative sentiments.

2.6 Opinion Mining Operation (OMO)

It can be said to be done in three main steps [19]:

- The first step would be to use the data mining methods. One may also make use of the natural language processing techniques so as to extract product features which have been suggested by customers, which we can name as **Determination of Subjectivity** [6]**.**

- The next step is to find an opinion sentence which is either positive or negative **(Determination of Polarity)** [10, 15].

- The last step is to identify the degree of the sentiment in terms of it being more positive, middle positive or weak positive, and similarly for a negative sentiment. Thus, the last step is to represent the **Strength of the Polarity.** After that, we type the conclusion of the result which has come up from the previous tasks [24].

Another parameter crucial for study on this topic is that of the subjective expression which can be used to express factual status. This status could pertain to the opinions, beliefs and emotions which are reflected in a text [7]. At the same time, an objective expression will show any information as per the intention of the writer. Opinion mining can recognise any subjective sentiment and can thus judge whether it is positive, neutral, or negative [7]. In addition, the process of Opinion will have in it natural language processing along with tasks such as text analytics. Furthermore, there is a natural linguistic processor which can split the text into sentences [7]. They also help in assigning tags, which include assigning categorisations such as noun, verb, and adjective to any given words. Here, the main aim is to analyse the sentiments in the given text [6].

At the same time, the researchers may use different techniques such as polarity tags, link-based patterns, document citations, semantic orientation, fuzzy pattern matching, stemming, punctuation, phrase patterns and stylistic measures. The sentiment analysis endeavours to determine the attitude of a speaker and relates it to the topic. It also relates the attitude to the overall contextual polarity present in the document [22].

This may be categorised as positive, negative or neutral. Moreover, the attitude can be deemed as "**beyond polarity**", which is an emotion classification rather than a classical sentiment analysis, not unlike emotional states such as *angry*, *sad* and *happy*. In this case, we label these as **multi-class classification** [23]. Here, the attitude could be the writer’s judgment, evaluation or the affective state in which the writer is typing statements or intending emotional communication which the writer wants to convey [24].

Some of the crucial research work present in the given subject is that of **Turney** [15] and **Pang** [10], who used many methods to judge the polarity of many reviews, including product and movie reviews.

In addition, there is a direction in research in which the ***subjectivity/objectivity identification*** of a text and this type of classification are to categorise the given sentence according to one of two classes (objective or subjective) [2].

Moreover, it can be seen from research by **Fangzhong Su et al.** [28] that the results will rely on the definition of subjectivity that is utilised when one is annotating texts. At the same time, **Pang** [2] talks of the removal of objective sentences from a document**.**

The same is suggested before the classification of the polarity. This could help improve performance. A more fine-grained analysis model, also known as ***feature/aspect-based sentiment analysis***, can explain another method of sentiment analysis. Here, one can determine the opinions expressed on different features which relate to entities such as cell phones, digital cameras, banks, etc. This aspect may be said to be a part of an entity, such as when the screen of a cell phone is being referred to, then the picture quality of a camera would be the attribute [29].

The detailed discussions of level of sentiment analysis are found in [30] **Liu's NLP** book in a chapter on *"Sentiment Analysis and Subjectivity"* (2010). Computers may also perform the automated sentiment analysis on digital texts. Such an analysis can be conducted with elements of machine learning, such as the *latent semantic analysis*, "bag of words," machines, and *Semantic Orientation* [15]. Additionally, some sophisticated methods can detect the holder of a sentiment. This refers to the person who has been in the affective state. The target may also be affirmed as the entity for which the effect is felt [31].

Sentiment Analysis can be thus divided into two separate categories: manual and automated sentiment analysis. Manual analysis refers to human sentiment analysis. The main differences between these two approaches are the efficiency of the system along with the accuracy of the analysis [32].

Some companies, such as Biz360 [33], have used a mix of the two methods. The Biz360 company uses machine learning to train the system through annotations which are created by Mechanical Turk, which additionally means that every item in a training set is annotated by a human being [33]. This may be carried out for better and accurate efficiency of the sentiment analysis. In the automated method, open source software tools will make use of statistics, machine learning and NLP techniques for a sentiment analysis [19].

Knowledge-based systems will make use of publicly available resources. This includes WordNet-Affect [34], SentiWordNet [35] and SenticNet [36] for the extraction of semantic and affective information related to natural language concepts.

Furthermore, one can see that a human analysis component is needed at this point of the sentiment analysis, such as the case of automated systems that are unable to analyse historical tendencies in an individual commenter [29]. Moreover, the platform could be classified incorrectly when there is an expressed sentiment. The impacts of the automation, which include 23 percent of comments, can be correctly classified by a human being [37].

It can be said that the structure of sentiments along with the topics can be very complex. There is also the problem of non-monotonic sentiment analysis in relation to sentence extension and stop-word substitution.

For example, there would be a difference in a tweet which says:

I will not let you shut down my mouth from speech.

They will not let you shut down my mouth from speech.

Thus, for this issue one needs **rule-based** and **reasoning-based** approaches to be practicing sentiment analysis [6]. This may be carried out by using **Defeasible Logic Programming** [38] which is a non-monotonic logic suggested by **Donald Nute** [39], which consists of three models: **strict rules** (the fact is always a consequence of another) which have priority during the process; **defeasible rules** (usually the fact is a consequence of another), which is applied if the priority of the defeater is lower than the defeasible rule priority; and finally, **undercutting defeaters** (exceptions will be specified to the defeasible rule) [39], [40]. One may also list a number of tree traversal rules which can be practiced in a syntactic parse tree for the extraction of the topicality of the sentiment presented in the open domain setting [41, 42].

The accuracy of a sentiment analysis system depends on human judgments and is measured by factors such as precision. The human raters feel that a 79-percent accurate program will do as well as it would be done by humans [33]. This accuracy, however, is not sufficient. In cases where a program was “right” 100 percent of the time, human beings would still disagree at the same rate, namely about 20 percent of the time. This is due to the fact that they may change and might disagree about their answers [33].

Thus, due to it being subjective, the evaluation of sentiment analysis systems is still very difficult. Sentiment analysis takes on returning a scale in comparison to binary judgement. At the same time, a correlation would be a better measure than precision as correlation will take into account any predicted value and its closeness to a target value [43]. The output in this case of opinion mining could be a ***feature-based opinion summary*** or an ***aspect-based opinion summary***. Here the sentiment classification could then be a sub-task. The present work is segregated into two main areas that show two kinds of opinions**,** one of which is mining regular, or what is known as, direct opinions.

For example, if a text says, *“**These shoes are great**”*, and simultaneously another text says, *“**After taking this medicine, I got a headache**.”* On the other hand, there is the mining of comparative opinions; an example would be *“**Maggi tastes better than Sun feast Noodles**.”* Here the analysis is complex as the sentence is a comparative statement and does not directly specify a positive or negative emotion [25].

2.7 Emotion Mining

Another process which has been used in this research and needs attention is emotion mining. Emotions are said to be the mental statement of the human mind and a psychological attribute. Thus, they are subjective and difficult to categorise.

We can see here the famous division of emotions by **Ekman** [44]**,** who divides the emotions in to six groups: happiness, disgust, anger, sadness, fear and surprise. Moreover, the emotion mining operation can be divided in to three practical gradients:

1- Determining the trend of the text as to whether it is positive or negative [10, 15].

2- The power of wealth which is offered by the text [6].

3- The measure of the level of excitement from the text [24].

One can list many methods to extract emotion from a text, some of which include:

- Keyword Spotting;
- Lexical Affinity measures; and
- Statistical Natural Language Processing techniques.

The first categorisation, Keyword Spotting, depends on a lexicon and a vocabulary. It also emphasises any emotional words in the text through the lexicon.

This technique is also seen as a smooth and quick technique. However, sentence structure may be very complicated and it will not be easy to detect emotions of the given text. One such example would be WordNet-Affect [34] which uses WordNet to conglomerate words into a number of groups called synonyms ("*synsets*").

Another example is the SentiWordNet [35], which splits WordNet synsets into two scales. These are divided into positive and negative scales. They may also be divided into subjective/objective scales. SentiWordNet is said to be a crucial tool for opinion mining as it can detect positive and negative emotions. Moreover, here the objective in a text would mean numerical scores [35].

The approach is that of measuring emotional weights which are possibly present in the natural language of a text that relies on keyword spotting. Thus for the emotional weight estimation, this approach would then separate the text into words and relationships which are present among the subject, verb and object [45].

SentiWordNet is required to have a large-scale database to test the accuracy of the given synset measures. Therefore, some approaches will become mixed when using both SentiWordNet and WordNet-Affect [46].

Some researchers have given more weight to WordNetSynset to obtain the different meanings of words as these same words may be used in many situations differently. They would convey different opinions. Moreover, the adjective and semantic orientation which is discovered currently in the text may be identified by using WordNet [47].

Furthermore, another technique which has been prevalent is the Lexical Affinity measure. This technique is the best of what has been seen in the first category. Here the process is such that each word will be assigned a probabilistic affinity. This will then pertain to definite feelings. For example, if the word is "***brave,***" then the probability of the word showing a positive feeling is about 90 percent. In the keyword spotting, the same will rely on a sentence corpus. However, it might not work with the same agility when analysing a complex text, such as the text "***Yazeed was not brave at all!***" Here, when one is measuring the probability of each emotion, it could depend on the text corpus which has been used in training.

The calculation measures every word which is then the percentage of emotional senses to the total senses which the word may have. In addition, WordNet-Affect and WordNet will look at the total number of senses and the number of emotional senses. This method is seen in many synsets of the word [48]. Furthermore, one may list here the contribution of those [48] who presented a model which synchronise the information measures along with the domain knowledge with the help of WordNet so as to extract concepts from the text.

The third technique employed is Statistical Natural Language Processing. This technique uses machine learning algorithms which analyse the relationship among lexicon affinity and the co-occurrence frequencies of the words [49, 50]. This is difficult to apply to social media due to the informal language along with the lack of structured sentences found in social media [5].

Another researcher defines a trustworthy and faster sentiment classifier and uses the existing tweets so as to classify new tweets [46, 5]. In case a new lexicon is not present in the polarity keywords, this new tweet is examined as per the adjective, verb, etc., which are the polarity keywords. Thus, through this approach, the classifier could reduce the development time of building. Also crucial here is the sentiment classifier, which can be used on the selected domain [5].

2.8 Sentiment Classification

There are many approaches to listing a sentiment. The major approaches, however, are the knowledge-based approach along with the machine learning approach [51].

The knowledge-based approach refers to the utilisation of affect dictionaries of opinion words which are defined in advance and seek input words [52].

Moreover, it analyses the effects of the input words. The machine learning approach simultaneously makes use of the statistical model which can be used to classify a sentiment in the form of input documents which are reliant on the vocabulary [53, 54].

CHAPTER 3

RESEARCH METHODOLOGY AND DATASETS

3.1 Introduction

In this chapter, we discuss our training corpora and test datasets which play an important role in our experiment along with the pre-processing and feature reduction operations that would be applied to the training corpora, the benefits of which have been gained from pre-processing and reduction of the feature space. Finally we describe the machine learning algorithm that has been used in extracting the experiment and in detecting the sentiment polarity of the tweets for the test dataset after applying them to the processed training corpora by using the MNNB supervised machine learning algorithm for text categorisation and by using the SMO machine learning algorithm.

3.2 Training Corpora

Our training corpora have been collected in advance from Twitter using API (Application Programming Interface) Twitter which was downloaded from the archive **http://help.sentiment140.com/for-students**. We then created groups which were collected sequentially and increasingly (i.e., creating groups by appending extra instances from the training corpora at the end of the previous one, which was also collected from the training corpora so as to obtain the following group and the same for the other groups in order to maintain the sequence when adding extra instances) from the training corpora to do our experiment. Pre-processing will be demonstrated in the next sections.

Moreover, the training corpora have been collected based on emotion icons such that the tweets containing emotion icons such as "**:)**" were positive and those with emotion icons such as "**:(**" were negative. As shown in Table 1, they demonstrate the queries which have been used for aggregating positive and negative tweets by reliance on positive and negative emoticons respectively. Therefore, we can notice here that the query is represented as a keyword search using the API search of Twitter.

Emoticons appointed to :)	Emoticons appointed to :(
:)	:(
:-)	:-(
:)	: (
:D	
=)	

Table 1 List of Emoticons Used for Fetching the Tweets

In the situation of the "**:(**" query, it returns "**:P**", which does not necessarily reflect a negative sentiment. Therefore, it is filtered from the Training Corpora.

Additionally, the corpora contain six fields as follows: polarity of the sentiment which includes three values (0, 2 and 4) for the negative, neutral and positive sequentially, tweet id which has been used for fetching the tweets, date of the tweet, tweet query, user name and finally the message or the body of the tweet.

We have overlooked four fields and taken into consideration only the polarity of the tweet and the tweet itself, which are needed to prove the research goals.

Thus our increasingly growing groups of tweets that have been collected sequentially from the main corpora are divided into six training datasets which are collected in even portion from positive and negative tweets. They begin as follows: 1,000 tweets, 5,000 tweets, 10,000 tweets, 25,000 tweets, 50,000 tweets and 131,042 tweets, along with their configurations that consist of pre-processing and feature space reduction.

Moreover, we would apply the stemming technique as well as remove any simplified stop words list over all groups in advance in order to prepare them for the machine learning algorithm to classify the sentiment of each tweet.

The researchers used the API search of Twitter to collect the tweets that built the training corpora depending on keyword search. These keywords or queries represent emoticons which the researchers used to fetch the tweets in advance as we mentioned previously.

3.3 API Twitter

As an interface for the Twitter social media application, and similarly for all operating systems which facilitate the reachability to the Twitter contents, it offers a search engine to extract tweets by depending on a keyword search. Furthermore, API Twitter represents a large part of the great success of Twitter due to its availability of data to developers and researchers around the world.

3.4 Analysis Levels

There are three novel approaches or three situations for sentiment analysis:

1. Classifying the sentiment at the document level (i.e. extracting the net sentiment that is obtained from the whole document).

2- Classifying the sentiment at the sentence level (i.e. extracting the sentiment gained one sentence at a time).

3- Classifying the sentiment at an attribute level which relies on specific factors such as topics, aspects, sentiment, opinion holder and the time of the tweet. This approach is known as *social listening*.

We have been using the second type of sentiment classification at the sentence analysis level to extract the sentiment of the whole tweet which mostly consists of more than one phrase along with its aspects [55].

3.5 Test Dataset

The test dataset is gathered manually and arbitrarily from multi-domains with specific queries. Products include (Twitter API, Kindle2), company (Google, McDonald’s), location (San Francisco, North Korea), person (Obama, Lebron), movie (Night at the Museum, Star Trek), event (India election, Googleio), miscellaneous (Stanford, dentist). These are obtained from 182 positive tweets and 177 negative tweets. Table 2 shows their distribution:

Domain	Number	Percentage over test dataset
Company	119 tweets	33.15%
Miscellaneous	67 tweets	18.66%
Person	65 tweets	18.11%
Product	63 tweets	17.55%
Movie	19 tweets	5.29%
Location	18 tweets	5.01%
Event	8 tweets	2.23%
Total	359 tweets	

Table 2 Distribution of Test Dataset Domains

The researchers removed the tweets that contained positive and negative emoticons in the same tweet during the fetching time of training the corpora, and ignored them in the case of manually gathering the test dataset in order to prevent overlapping. They marked the positive features as a part of negative features and vice versa.

We consider this example tweet: "***I'm tired today :(but my birthday is tomorrow :)***". The two emoticons in the this tweet will cause the machine to be overlapped when deciding the correct trend for the text so that such tweets will be removed.

3.6 Pre-processing (P.P), Feature Reduction (F.R) and Feature Selection (F.S)

In this section, we will discuss our steps to process the original tweets from the training corpora and to extract as much as possible of the pure meaning of the text in order to conduct our experiment in classifying the sentiment of tweets. We will first determine the major benefits from these steps as follows:

- Reduce the dimension of the feature space without losing the anticipated meaning of the tweets.
- Remove unwanted features or irrelevant and useless features.
- Improve the performance and accuracy of the machine learning algorithm used for sentiment classification.
- Reduce training time, which is noticeable in large training datasets.
- Allow us to increase the size of a training dataset rather than an unprocessed one. [56, 57]**.**

3.7 The Steps of Pre-processing and Feature Reduction and Selection

1- Removing Target names and replacing them with an equivalence class: **TNAME**, Such as**:**

"**@GVDesign yes but I did not ask for it.**" Will be:

"**TNAME yes but I did not ask for it.**"

Therefore, the @ sign in Twitter is employed to call usernames in tweets. When a username is preceded with the @ sign, it will be a link to the profile of the user account on Twitter [58].

2- Removing the Uniform Resource Locater (**URL**) that is commonly used for Hypertext Transfer Protocol (**HTTP)** and Secured Hypertext Transfer Protocol (**HTTPS**) Internet links. Other links such as "**MAILTO**" and File Transfer Protocol (**FTP)**, and replace them with an equivalence class "**URL**" as shown in this example:

> "**Some critique would be good guys http://tinyurl.com/cq52kc**" will be:
> "**Some critique would be good guys URL**".

3- Removing hash topic tags and replacing them with an equivalence class **HTAG**; for example:

> "**@Widgetty Thanks for #followfriday yesterday**" will be:
>
> "**TNAME Thanks for HTAG yesterday**"

The hash tag represents a whitespace word preceded by the symbol **#**. Thus the hash symbol is frequently used in information technology to highlight the special meaning of some words to indicate it as a special topic (trend topic). It is widely used in messages of microblogging social web sites such as Twitter, Facebook, and Instagram [58].

The hash tags are helpful in detecting messages and grouping them. It may be considered to be an essential keyword that categorises documents or messages according to their meaning as to be relevant to a product, movie, person, news, companies domain and depending on their appearance in target documents or messages that one would wish to categorise [58].

4- We remove duplicate letters from the words that have more than double appearances of the same letter in the same word sequentially and substitute them with just one duplication, because some words have in the original shape and one duplication of a letter in words such as **feel, sell, surround, speed, kill** and so on.

As is common knowledge, most social media pioneers and Twitter pioneers use emotion words and shape them with duplications of letters to deeply express satisfaction or resentment about events, products, movies and so on. Therefore, they are used this way and in other ways to reflect their real feelings. The following are some real examples from our training corpora:

"**I loooove the Brady Bunch in all of its corny goodness!** "

"**I don't feeeel good I wanna go home.**"

Thus and in order, the two sentences will be:

"**I loove the Brady Bunch in all of its corny goodness!** "

5- Pick up the emoticons outside the training Corpora and considered them as noisy labels, and give the classifier an opportunity to learn from other features such as unigram, bigram and trigram.

Some tweets may contain more than one emoticon in their body text. There are three situations, one of which being that the user attempts to express more than one event or interactivity feelings through only one sentence.

Of course without consideration to the developers and researchers who try to gain benefits by taking samples from the user's tweets to analyse and detect some information that is useful in most fields of life. Therefore, the life fields that may gain benefits from this area of study may include study of the human brain and how to think, study of the human psychology of after the recent changes occurring around the world, checking whether or not some products and companies are satisfied with their customers to enhance their quality and quantity, and so on for other domains such as movies, television channels, companies, sports, news, websites, etc.

The second situation may concern some users who are ignorant of how to use and express their feelings towards something in a perfect way. Additionally, the use of slang language is not recognised even by the acronym dictionary. Finally, some users do so for entertainment and do not have any explicit idea how to express it. Some examples show the overlapping that may occur with this approach and yield the unacceptable accuracy during classification using the machine learning algorithm. For example:

"**Imam Hussein suns of Fatima :), but goddamn Yazeed kills him wrongfully**"**.**

"**I love Fatima daughter of messenger Mohammed :(".**

"**I am lucky because I will graduate from Cankaya University after some days :(".**

We see here the confusion, when they are kept in the body of tweets as the texts of the tweets discuss something good or bad in spite of the emoticons referring to an opposite meaning. This represents one of the limitations in our research.

6- The remove stop word list was inspired by Google and some extra cases were added by the researcher in such a way so as not to affect the meaning of the sentences, such as excluding negation elements: "**don't**, **shouldn't, isn't**, **didn't**". This will be illustrated with examples in the results chapter.

7- We use the acronym dictionary to translate the most famous abbreviations and acronyms to their polarity of sentiment by including them in the process steps of the MNNB and SMO machine learning algorithms.

Therefore, it is satisfying to find a suitable analysis trail. Twitter text classification looks at a varied set of short-text messages containing abbreviations and slang, as illustrated in Table 3.

Acronym	**Meaning**
LYL	Love You Lots/ Positive
OMG/omg	Oh My God/oh my god/ Positive or Negative
bff	best friend forever
uh uh	Negative
bah	Negative

Table 3 Samples of Acronyms

8- The use of word stemming to retrieve derived words to their origins using the Iterated Lovins stemmer.

We will remain here to demonstrate stemming, how it works, how it can be harnessed to serve our approach, as well as to explain the Iterated Lovins stemmer, when and in which states it is useful, and finally some examples about stemmer processing.

3.8 Stemming Algorithm

The stemming algorithm is a set of rules which are responsible for retrieving or reducing words to their original forms by removing the derivations and inflection suffixes from the word. All steps will be carried out as computational procedures. It is used heavily in areas such as IR (Information Retrieval) and CL (Computational Linguistics) [59].

3.9 Iterated Lovins stemmer

The iterated Lovins stemmer algorithm is considered to be the first published stemming algorithm and it contains a broad range of endings, conditions and transformation rules. Thus, we can conclude from its name (Iterated) to be representative of recursive procedures that begins to strip out the ending strings respectively from the end of a word towards its beginning.

Here two examples are mentioned to explain the workings of the Iterated Lovins stemmer: **relatedness, willingness** [59], as shown in Table 4.

Derivative and inflective word	First step	Second step	root word
relatedness	remove –ness	remove -ed	relat
willingness	remove –ness	remove -ing	will

Table 4 Iterated Lovins Stemmer Work

3.10 Machine Learning Algorithm

Machine learning algorithms can discover how to achieve significant tasks by popularising from examples. This is for the most part practical and cost-effective. Additionally, the recent availability of data addresses more problems to be tackled. Machine learning is used frequently and widely in fields which depend on knowledge extraction. The learning is a combination of three components [60]:

- **Representation:** Meaning the classifier must be formed using formal language which must be understandable and can be processed by a computer. Therefore, if the space of the problem is not matched with capabilities of the classifier, it will lead to *"the classifier cannot be learned."*

- **Evaluation:** This stage of learning distinguishes the good classifiers from the bad ones and is thus called the "scoring function".

- **Optimisation:** Here, and in the final stage, is important to highlight the optimum classifiers among the good classifiers which have been extracted in the previous stage.

Learning algorithms may be classified into three major types: Unsupervised, Supervised and Reinforcement learning algorithms. Additionally, there is a derived type resulting from supervised learning due to some restrictions, which is referred to as the Semi-Supervised learning algorithm. The differences between them are illustrated in the next sections.

3.11 Unsupervised Learning Algorithms

Here, the training dataset is not required. It directly and simply gives the output from the incoming data. Moreover, it is easy and rapidly processes the input data when implementing the task; however, it has limits to its accuracy due to the absence of the relation with respect to the taken samples [61].

3.12 Supervised Learning Algorithms

Supervised learning refers to when the algorithm is not specified along with its structure. Furthermore, some of the parameters are still anonymous. Therefore, we need to provide a training dataset to identify the structure of the algorithm along with the unknown parameters. So for testing the dataset, the algorithm will provide the expectations relevant to the input data and the parameters that have been learned from the training dataset.

Therefore, supervised learning has more flexibility than unsupervised learning, and is also more applicable in the case of predicting the incoming test dataset on the similar assumption distribution that has been shared between testing samples and training samples [61].

3.13 Semi-Supervised Learning Algorithms

Semi-supervised learning is a complement to Supervised learning to solve problems that may be encountered by supervised learning algorithms. Some real applications do not have ability to provide sufficient labelled training data, so supervised algorithms here will create the pitfall to the desired accuracy due to the training set not being sufficiently large.

In addition to the small training dataset, the high dimensionality of the data samples also creates restrictions and a lack of high performance quality of the supervised learning. Therefore, semi-supervised learning will allow the unlabelled data to be available in the training dataset. The semi-supervised learning will produce a reasonable estimation in the case of labels that are rare in the training dataset [61].

3.14 Reinforcement Learning Algorithm

This type of machine learning refers to suggestions and procedures about *what to do and how to plan* in order to maximize a numerical reward signal. Here, the learning does not mean which procedures or actions must be taken to solve the problem because other kinds of machine learning alternatively discover, by trying them, which procedures will generate the most rewards to solve the problem.

Therefore, reinforcement learning can be specified by describing the parameters and characteristics of the learning problem rather than determining the characterisations of the learning methods. In this case, we choose any learning method that is more suitable to solve the problem and will be considered as a reinforcement learning method [62].

3.15 Multi-nominal Naïve-Bayes for Text Categorisation

It is classified under supervised learning algorithms which are applied in the practical application of this dissertation. Naïve-Bayes is mostly used for text classification problems because of its high efficiency and smoothness when applied computationally in problems such as those that deal with text categorisation. We now demonstrate the computational operation for the class of a given tweet and so we declare some parameters in advance as follows:

$\boldsymbol{C}$ represents a set of classes, $\boldsymbol{N}$ *is* the size of the tweet (number of the words per tweet), after which the MNB assigns a test tweet $\boldsymbol{t_i}$ to the class with the highest probability $\mathbf{P}(\boldsymbol{c}|\boldsymbol{t_i})$, and by using Bayes' rule thus:

So $\mathbf{P}(\boldsymbol{c})$ which represents a prior class can be evaluated by dividing the number of tweets that are assigned to class c by the total number of tweets. $\mathbf{P}(\boldsymbol{t_i}|\boldsymbol{c})$ is the probability of obtaining a tweet $\boldsymbol{t_i}$ in class $\boldsymbol{c}$, which is calculated as:

Here we have the parameter $\boldsymbol{f_{ni}}$ as a counter of word $\boldsymbol{n}$ in the test tweet $\boldsymbol{t_i}$ and $\mathbf{P}(\boldsymbol{w_n}|\boldsymbol{c})$ as a probability of word $\boldsymbol{n}$ belonging to class $\boldsymbol{c}$**.** Furthermore, the probability that is estimated from the training tweets represents the last probability as follows:

where the count of word $\boldsymbol{x}$ in all training tweets that is given class $\boldsymbol{c}$ is represented by the parameter $\boldsymbol{F_{xc}}$, and by using the Laplace estimator to prime the count of each word with one to prevent the zero-frequency problem. Now we have to compute the normalisation factor **P(*t*)** in equation 1 as:

We can notice in equation 2 that there is a computational expense terms,

which can be ignored without any effect on the expected results, due to no dependency on the class ***c***, so the equation, will be modified to be written as follows:

Where $\boldsymbol{\alpha}$ represents a constant which will drop out due to the normalisation step [63].

3.16 Sequential Minimal Optimisation Learning Algorithm

This algorithm was invented by **John Platt** in 1998 to solve the quadratic programming problem which focuses the Support Vector Machine learning algorithm during training [64]. The quadratic programming problem is a type of mathematical optimisation problem. The optimisation problem means maximisation or minimisation of a real function; for example:

Suppose we have a function such as f: $Z \longrightarrow R$, from some set Z to R (Real numbers).

We want to seek element y_0 in set Z so that $f(y_0) \leq f(y)$ for all y in Z. This represents minimisation. If $f(y_0) \geq f(y)$ for all y in Z, this represents maximisation.

Moreover, it considered to be a supervised machine learning algorithm which can deal with nominal class, missing class values, binary class, as well as treat unary, nominal, numeric, missing values, binary and empty nominal attributes.

However, it cannot manipulate string attributes directly unless it converts them to another type, such as numeric attributes. and the class to be as nominal, Only in the case of using a special function known as the StringKernel function, the SMO can treat string attributes directly without conversion to another type, but it still gains low accuracy in contrast to using another kernel function, such as the Polynomial Kernel function. Furthermore, the Polynomial Kernel function needs the user to convert string attributes to the numeric type in order to accept the accuracy of the sentiment classification when applying the test dataset on the training corpora.

CHAPTER 4

FINDINGS, RESULTS and DISCUSSION

4.1 Introduction

In this chapter, we discuss the findings of the research, assumptions and the results obtained by using the methodologies demonstrated in the previous chapter, namely using MNNB and SMO machine learning algorithms with the ability to analyse the texts of tweet microblogs to detect emotions mentioned inside them. These tweets are collected from multi-domain training corpora rather than only a specific domain. Moreover, the classification of tweets using machine-learning algorithms relies on factors such as word presence, word frequency, stemming of words, simplified stop words list, converting tweets to lower-case letters and an acronym dictionary and by taking into consideration the situations of the mentioned factors.

Moreover, the thesis discusses the pre-processing operations, features reduction and makes comparisons among the increasingly growing groups which have been collected from the main corpora according to the results that have been gained after these actions.

Here we prove the objectives which were obtained from the achieved results of the research. One of them says, word presence is not a default factor to obtain high accuracy of classification in contrast to other factors using the unigram feature extractor. Therefore, we precisely mean the word frequency because in some cases, the reverse state is seen and the high results were obtained by taking into account word frequency as a factor of extraction in the classification of tweets using the MNNB and SMO models.

Furthermore, another important objective is that of groups of tweets increasingly yielding high results incrementally in the case of the bigram and trigram feature extractors.

We demonstrate the pre-processing, feature extraction and reduction steps using a block diagram that shows all the steps briefly with explicit distribution of the work among units until the final step of the emotion detection class for the input tweets is reached.

It is important to note that this diagram is not unique. Some parts of the diagram have the same names as others. Furthermore, these parts are responsible for the transections and may be repeated frequently around the world of sentiment processing steps. However, new experiments will occur and the results obtained from these experiments will be discussed. (See Figure 4)

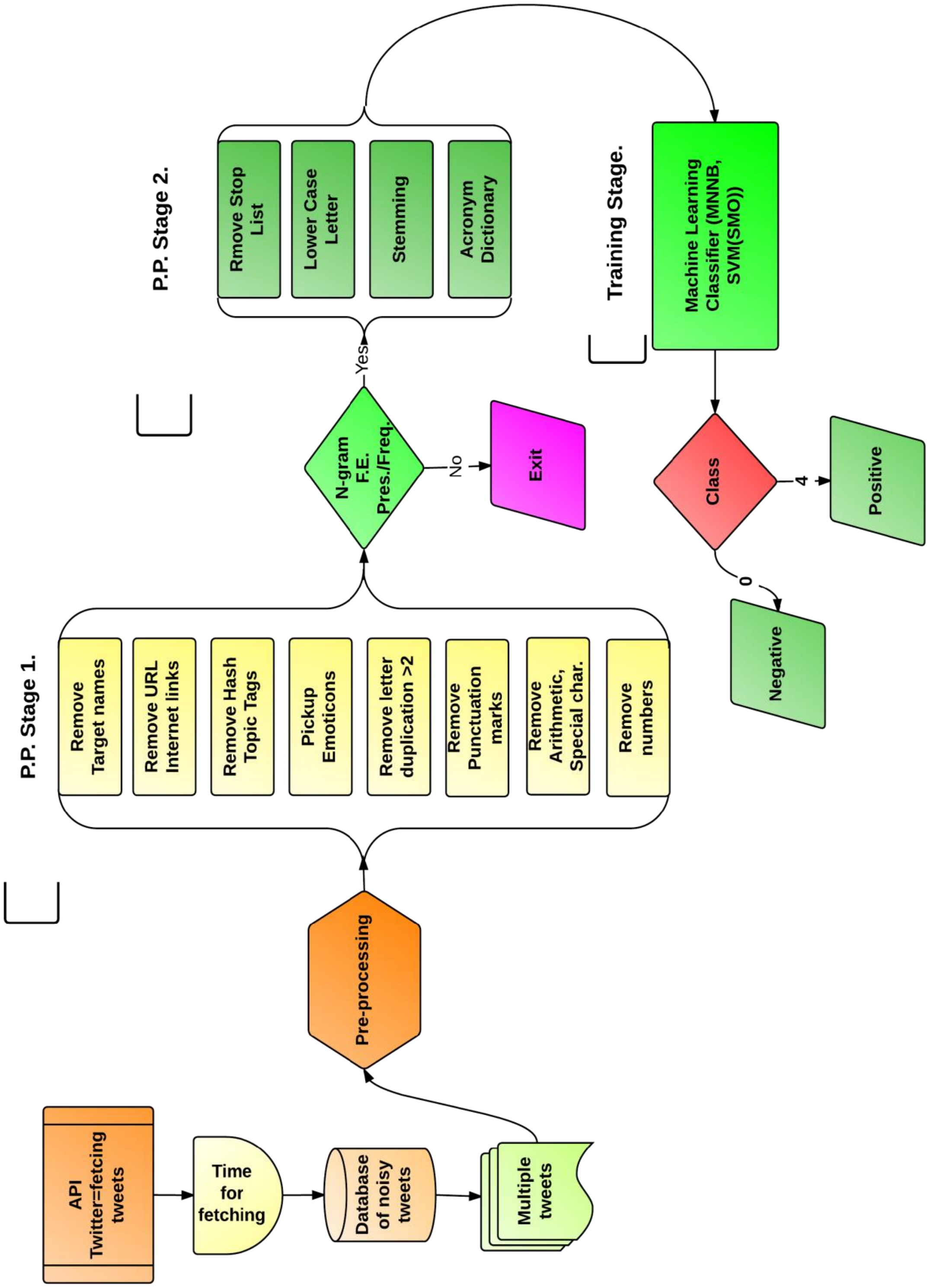


Figure 4 Flow Chart of emotion detection steps

4.2 Pre-processing and Feature reduction Results

After applying the pre-processing steps to the increasingly growing groups which were gathered from the training corpora, as well as applying the pre-processing to the test dataset, one can note the reduction of the feature space by selecting only the effective features and ignoring the unwanted features which do not have an active presence in the classification using the supervised machine learning algorithm. Note that @user name was removed in the preparation step of noisy tweets, so we have here only @target names which will be replaced with **TNAME** as mentioned and explained in the pre-process steps in Chapter 3. (See table 5)

Groups	**1000**	**5000**	**10000**	**25000**	**50000**	**131042**	**359 test**
P.P. steps	**tweets**	**tweets**	**tweets**	**tweets**	**tweets**	**tweets**	**dataset**
T.N.	530	2511	4964	12228	24135	63086	106
URL	43	226	401	1115	2045	5538	58
H.T.T.	18	82	168	462	1077	2676	32
Punc. Marks	1842	10182	17750	44619	89423	236265	939
Arith., spec.	122	425	866	2145	4253	14032	66
Numbers	288	2225	3066	7742	15335	39884	145
Duplicate letters > 2	89 case	424 case	903 case	2108 case	4314 case	11660 case	19 case

Table 5 P.P. Results of Increasingly Growing Groups of Corpora

After the pre-process steps, the differences between the word vector space before and after pre-processing and feature reduction operations can be seen in Table 6.

Groups Actions	1000 tweets	5000 tweets	10000 tweets	25000 tweets	50000 tweets	131042 tweets	359 test dataset
Before P.P	4167	13187	22020	42486	69129	138839	2025
After P.P	3514	10263	16788	29350	44430	82628	1785
After P.P+S.L	2968	9433	15856	28273	43245	81271	1373
After P.P+L.C	3017	8508	13825	23607	34979	64602	1480
After P.P+Stm	2364	5968	9251	15395	22166	41408	1253
After all	2103	5564	8742	14727	21359	40385	1040

Table 6 Feature Space Before and After the P.P. and F.R. Steps

Another process which has been used here and which needs attention is the stop words list. In this research, we proposed a simplified stop words list in which no negation is removed from the corpora so as to keep the detection of negative emotions which mostly have the negation words (such as no and not etc.) in the expressions, and the same for most of the adverbs so as to keep important connections between words in order that they remain meaningful. For example:

"**Ali did not eat good meal**". Suppose here, we use the trigram feature extractor to detect the sentiment of the sentence, as shown in Table 7.

Ali did not	did not eat	not eat good	eat good meal
Negative	Negative	Negative	Positive

Table 7 Extract Features Using Trigram

It can been seen that the polarity of the negation (3 times) is greater than the polarity of the positive (1 time), so this sentence will be classified under the negative polarity, while removing the default stop words list from our training corpora will give us the contrasted meaning of the reality of the sentence, and will be as follows after picking up the default stop words.

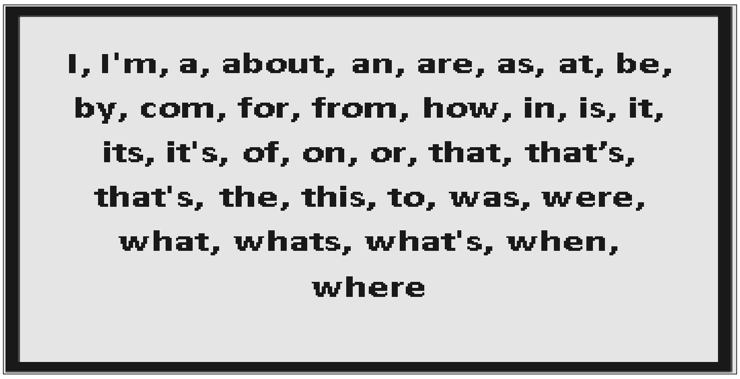
"**Ali eat good meal.**", and Table 8 shows that:

Ali eat good	**eat good meal**
Positive	Positive

Table 8 Pickup Default Stop Words from Trigram Feature Extractor

Therefore, the real meaning of the sentence is lost, which will categorise it under the positive polarity. As a result, the simplified stop word list was used to prevent this loss of meaning from occurring.

Figure 5 below shows our stop words list which was inspired by what is believed to be the Google stop words a decade ago. Therefore, some extra cases which do not affect the meaning of the sentences, such as negation stop words, were added.



I, I'm, a, about, an, are, as, at, be,
by, com, for, from, how, in, is, it,
its, it's, of, on, or, that, that’s,
that's, the, this, to, was, were,
what, whats, what's, when,
where

Figure 5 Stop words used

4.3 Supplying Test Dataset on Training Corpora

Now and after training and learning our increasingly growing groups of corpora and the test dataset through the machine learning algorithms MNNB and SMO, we will view the best results obtained from supplying test dataset on increasingly growing groups of training corpora and in dependence on important factors such as bag of words (word presence or word frequency), simplified stop words list, iterated Lovins stemmer and lower-case letters. Furthermore, this research offers a Chapter of Appendix of all tables.

Now, Table 9 shows the final results obtained from using unigram the MNNB sentiment classifier on all groups. Most notably, all increasingly growing groups show an equal distribution between positive and negative tweets. Moreover, they are collected sequentially or randomly from the main corpora.

Table 9 summarises the highest results which were obtained from the unigram MNNB model and for all groups. See unigram MNNB tables in Chapter 6 (Appendix of Tables (Tables 1 to 12)) and note some of the highlighted rows of tables which represent the situation of obtaining high accuracies.

Group	Feature	S.L.	Stm.	L.C.	Pres./Freq.	Accuracy
1000 tweets	unigram	removed	no/yes	no	Pres./Freq.	69.0808%
5000 tweets	unigram	removed	no	yes	Pres.	75.2089%
10000 tweets	unigram	removed	no	no	Pres.	72.1448%
25000 tweets	unigram	removed	no	no	Pres.	74.6518%
50000 tweets	unigram	removed	no	yes	Pres.	76.6017%
131042 tweets	unigram	removed	no	yes	Freq.	76.8802%

Table 9 Classification Accuracy of MNNB Using Unigram

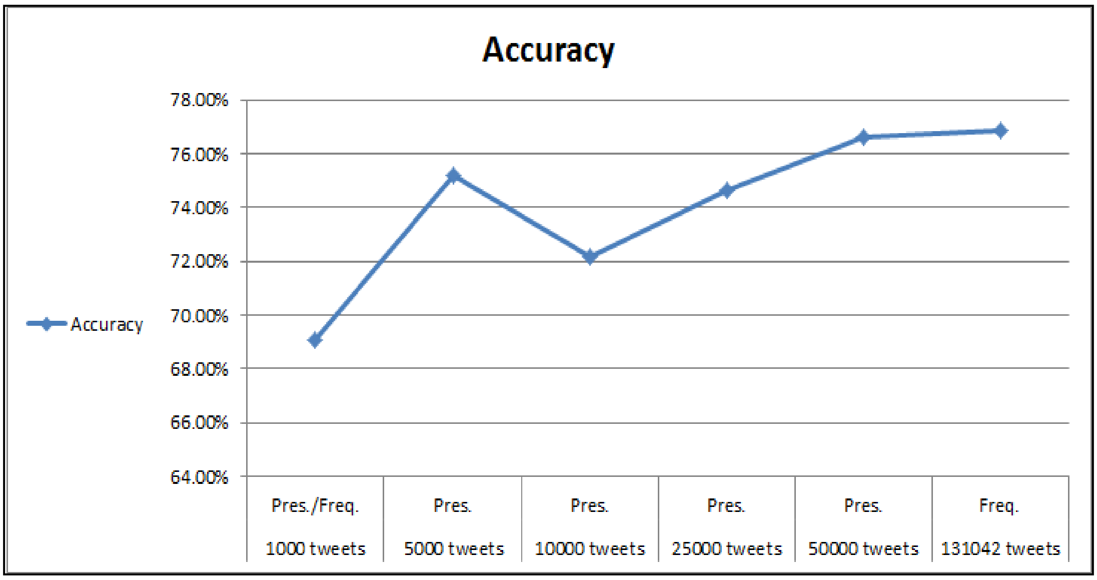


Figure 6 MNNB results using unigram feature extractor

It is interesting to note that high accuracies were obtained using word presence as a factor of extraction in the case of word tokenisation (as confirmed by **B. Pang et al.** [9] who depended on the training dataset collected from movie review domains (single-domain)). Moreover, high accuracies through reliance on word frequency as a factor of extraction were also obtained, but this occurred with the training corpora collected from multi-domains. This can be seen in the first and last group. (See Table 9).

The same will be done for the bigram and trigram feature extractors, as shown in Tables 10 and 11.

Note: Table 10 summarizes the highest results which were obtained from the bigram MNNB model and for all groups. See the bigram MNNB tables in Chapter 6 (Appendix of Tables (Tables 13 to 24)) and note for some highlighted rows of tables which represent the situation of obtaining high accuracies. The same situation occurs for Table 11 which also summarises the highest obtained results but in the case of the trigram MNNB model. For more details see Chapter 6 (Tables 25 to 36).

Group	**Feature**	**S.L.**	**Stm.**	**L.C.**	**Accuracy**
1000 tweets	bigram	removed	no	no	70.195%
5000 tweets	bigram	removed	no	yes	74.3733%
10000 tweets	bigram	removed	no	no	74.9304%
25000 tweets	bigram	removed	no	yes	76.3231%
50000 tweets	bigram	removed	yes	yes	76.3231%
131042 tweets	bigram	removed	no	yes	77.1588%

Table 10 Classification Accuracy of MNNB Using Bigram

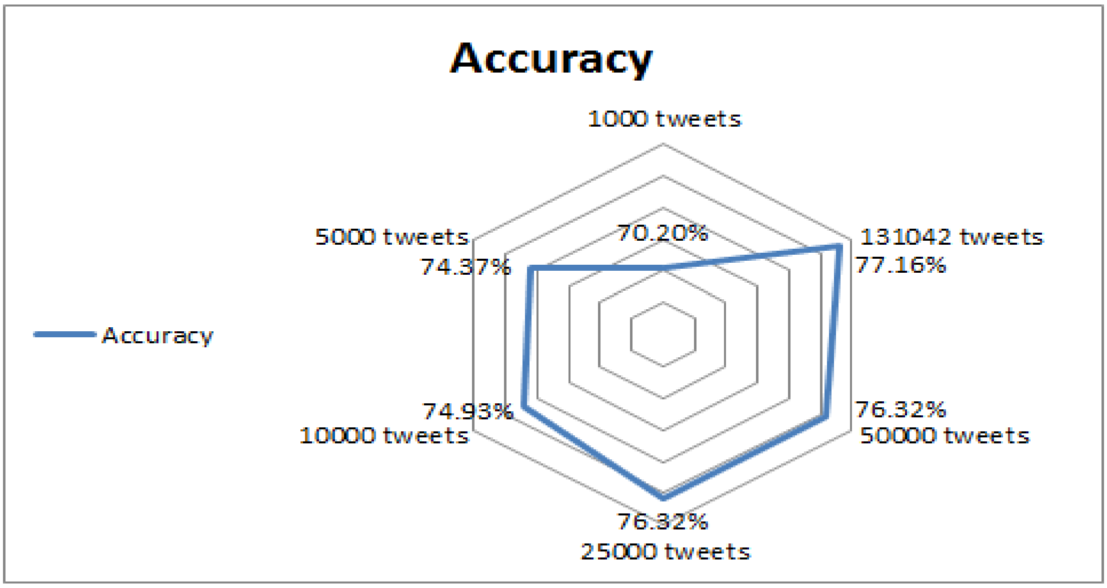


Figure 7 MNNB results using bigram feature extractor

Group	Feature	S.L.	Stm.	L.C.	Accuracy
1000 tweets	trigram	removed	no	no	70.195%
5000 tweets	trigram	removed	no	no	73.8162%
10000 tweets	trigram	removed	no	no	73.5376%
25000 tweets	trigram	removed	no	yes	74.9304%
50000 tweets	trigram	removed	yes	yes	76.0446%
131042 tweets	trigram	removed	no	yes	76.8802%

Table 11 Classification Accuracy of MNNB Using Trigram

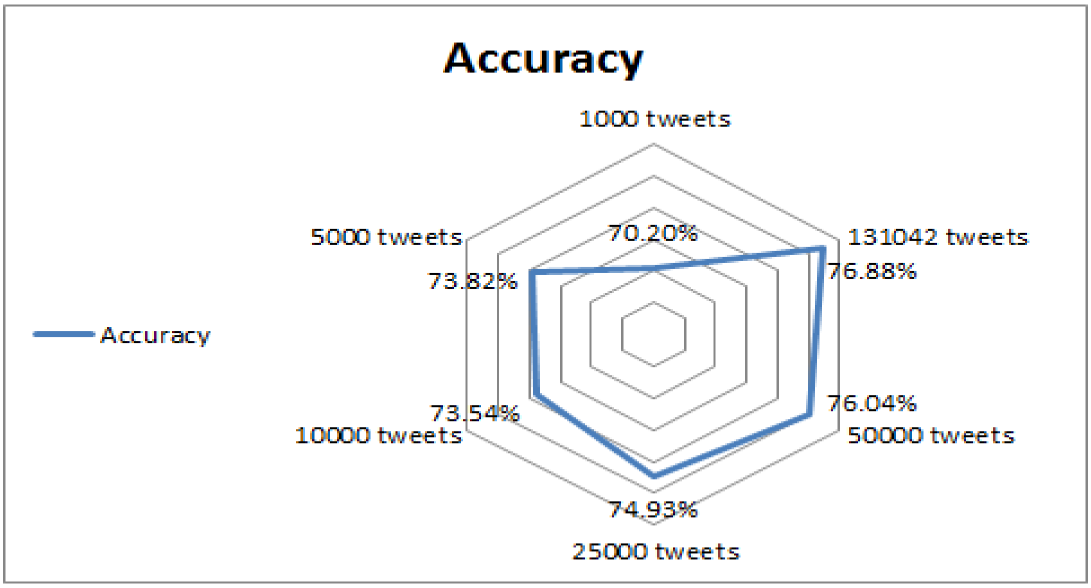


Figure 8 MNNB results using trigram feature extractor

Another important finding was that as long as the number of instances of tweets increased, the accuracy of sentiment classification also increased, which is clear, especially in the case of using bigram and trigram feature extractors. Except for two atypical states occurring between 25,000 tweets and 50,000 tweets in the case of using the bigram feature extractor, the accuracy remains unchanged, and between 5,000 tweets and 10,000 tweets with a difference of only 0.28 for the 5,000 tweets exceeding 10,000 tweets in the case of using the trigram feature extractor. However, as an accumulative result, the accuracy increased incrementally and explicitly with increases in the number of instances of tweets. For example, in the case of the bigram feature extractor, the accuracy beginning from 1,000 tweets to 131,042 tweets have increased about 7 degrees.

Also we have other experiments which were obtained from supplying a test dataset on the increasingly growing groups (1,000 tweets, 5,000 tweets, 10,000 tweets and 25,000 tweets), by using SMO classifier model and through depend on the same feature extractors as in the case of MNNB model. The results are shown in the Tables 12, 13, 14.

However, it is obvious for the all, after checking the results, the MNNB algorithm exceeds the accuracy of other algorithms, and SMO model was an example to prove that. The measurement of the efficiency includes time consumption, high accuracy of classification and the number of tweets that one can train and test with the two models.

It is important to understand that the SMO does not work directly with string attributes. Therefore, the string attribute will be converted to numeric attributes so as to be accepted by the algorithm. Moreover, all the obtained results from the SMO model depend on a specific number of words per class which was here just 1,000 words to keep the assigned class.

For example: in the case of the unigram feature extractor if the word account for the training group (10,000 tweets) were changed from 1,000 words per assigned class to 14,000 words, more unwanted attributes may be acquired. Therefore, the feature space will be increased and cannot be manipulated easily nor will it take more than the expected time to be processed. The number of attributes will also increase from 1,022 attributes to 13,478, which will lead to a decrease in the accuracy of classification from 73.8162% to 71.0306%.

Group	Feature	S.L.	Stm.	L.C.	Pres./Freq.	Accuracy
1000 tweets	unigram	removed	no	yes	Pres.	67.688%
5000 tweets	unigram	removed	no	yes	Freq.	71.3092%
10000 tweets	unigram	removed	no	yes	Pres./Freq.	73.8162%
25000 tweets	unigram	removed	no	yes	Pres.	74.0947%

Table 12 Classification Accuracy of SMO Using Unigram

Note: Table 12 summarises the highest results which were obtained from the unigram MNNB model for all groups. See unigram MNNB tables in Chapter 6 (Appendix of Tables (Table 37)) and note some of the highlighted rows of the table which represent obtaining high accuracies.

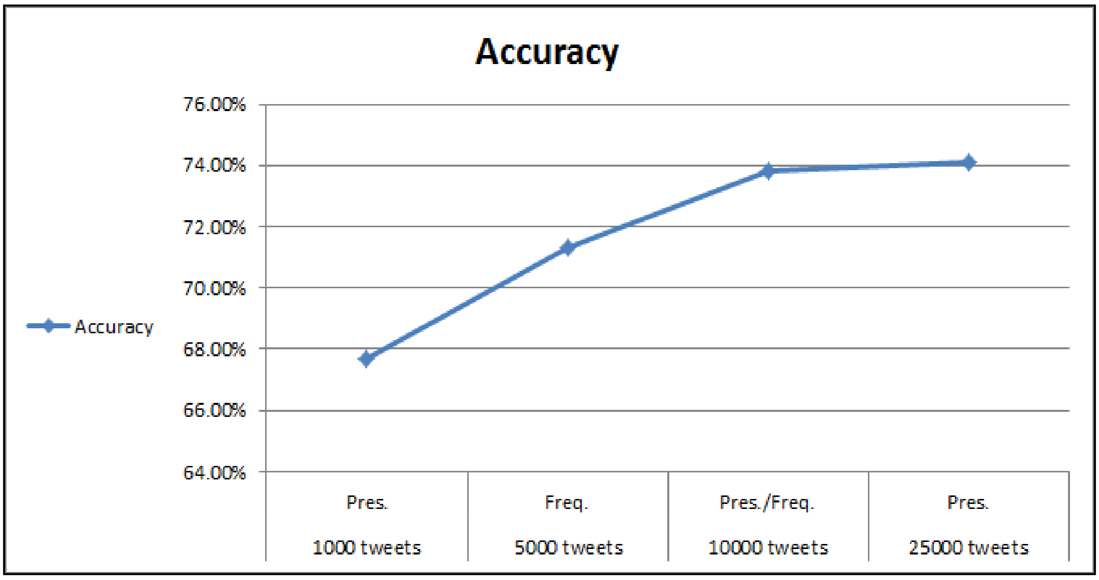


Figure 9 SMO results using unigram feature extractor

Group	Feature	S.L.	Stm.	L.C.	Accuracy
1000 tweets	bigram	removed	no	yes	66.0167%
5000 tweets	bigram	removed	no	yes	68.5237%
10000 tweets	bigram	removed	no	yes	72.9805%
25000 tweets	bigram	removed	no	yes	74.0947%

Table 13 Classification Accuracy of SMO Using Bigram

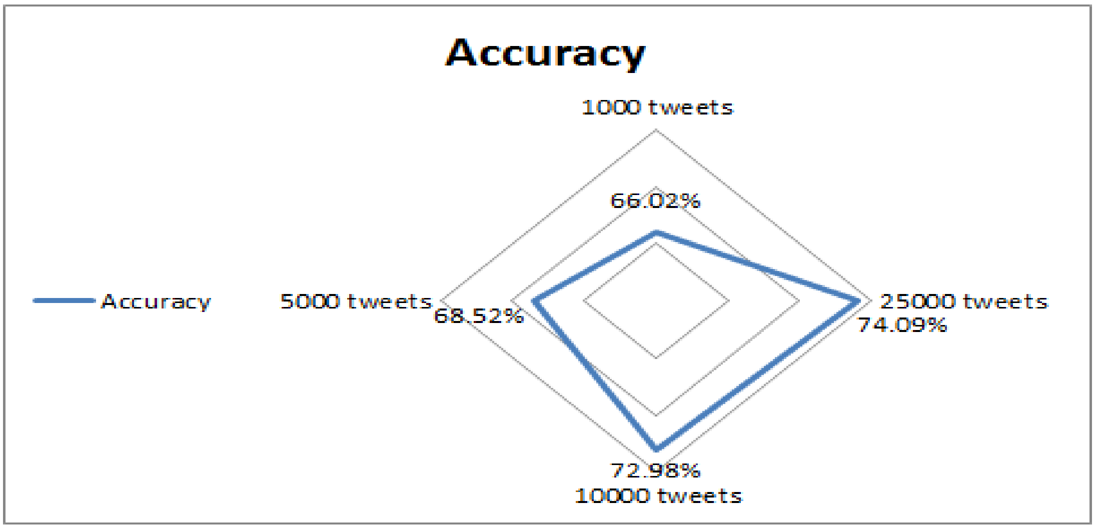


Figure 10 SMO results using bigram feature extractor

Group	Feature	S.L.	Stm.	L.C.	Accuracy
1000 tweets	trigram	removed	no	yes	64.624%
5000 tweets	trigram	removed	no	yes	67.688%
10000 tweets	trigram	removed	no	yes	72.7019%
25000 tweets	trigram	removed	no	yes	72.7019%

Table 14 Classification Accuracy of SMO Using Trigram

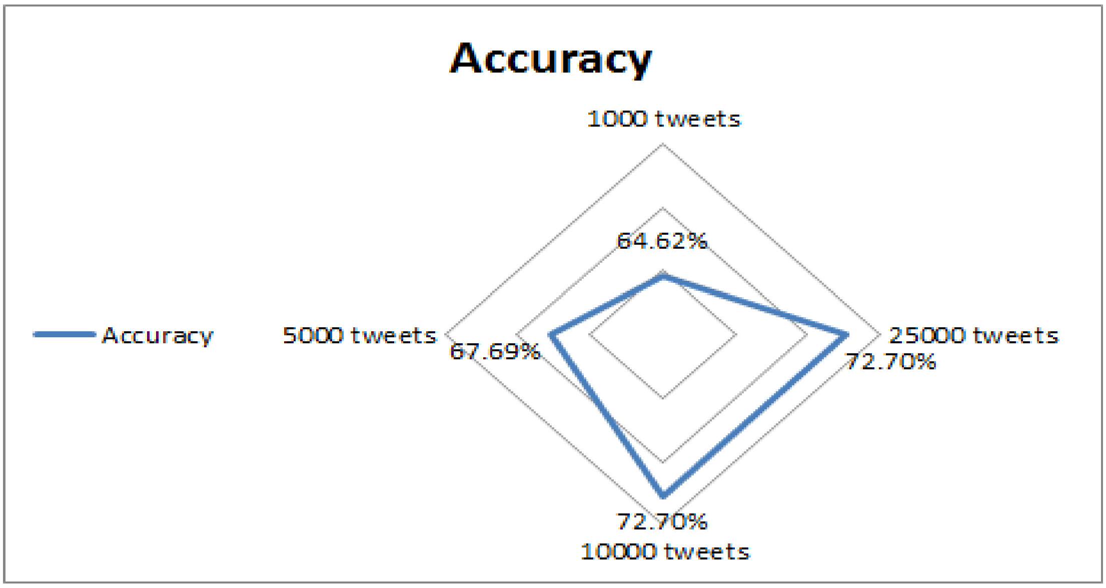


Figure 11 SMO results using trigram feature extractor

As the MNNB model, the SMO satisfies the first and second objectives, so long as they obtain high accuracy using word presence and in the case of the unigram feature extractor, the word frequency also obtained high accuracy. This is obvious in the 5,000 tweets group and 10,000 tweets groups (Figure 9). Moreover, it satisfies the concept of increasingly growing groups leading to obtaining high accuracies incrementally from the bigram and trigram feature extractors (see Figures 10 and 11).

In addition, we satisfy the second objective by obtaining high accuracies incrementally in the case of the unigram feature extractor without any atypical states, as explained in Table 12 and Figure 9.

Figures 12, 13 and 14 below demonstrate how the MNNB classifier model outperforms the SMO classifier model through the experiments obtained from the fourth selected increasingly growing groups of 1,000, 5,000, 10,000 and 25,000 tweets.

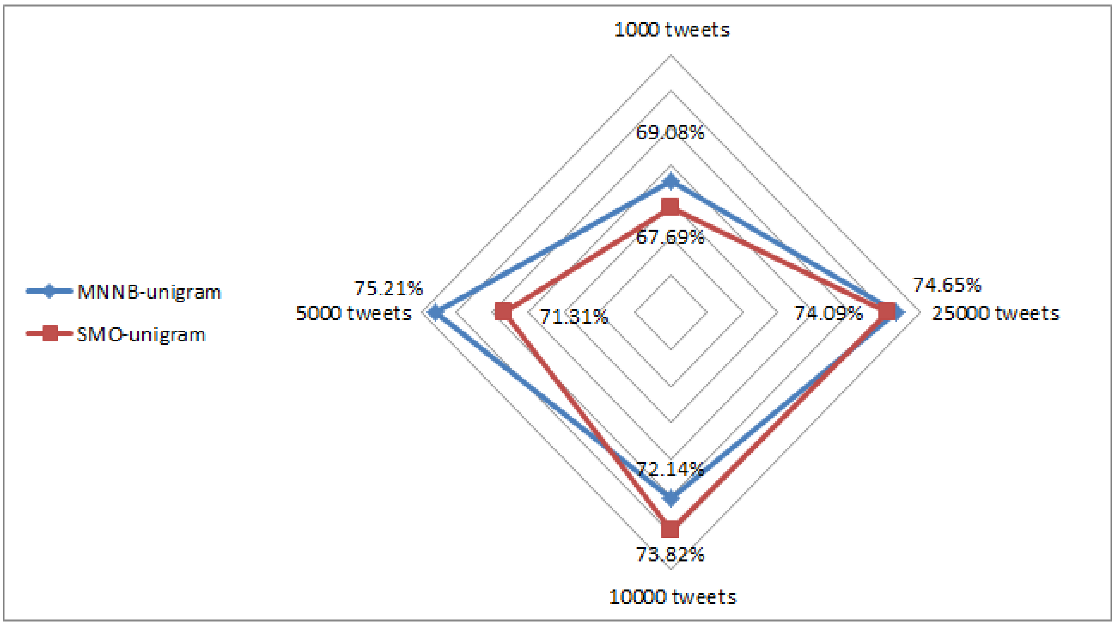


Figure 12 MNNB and SMO classification accuracy using unigram

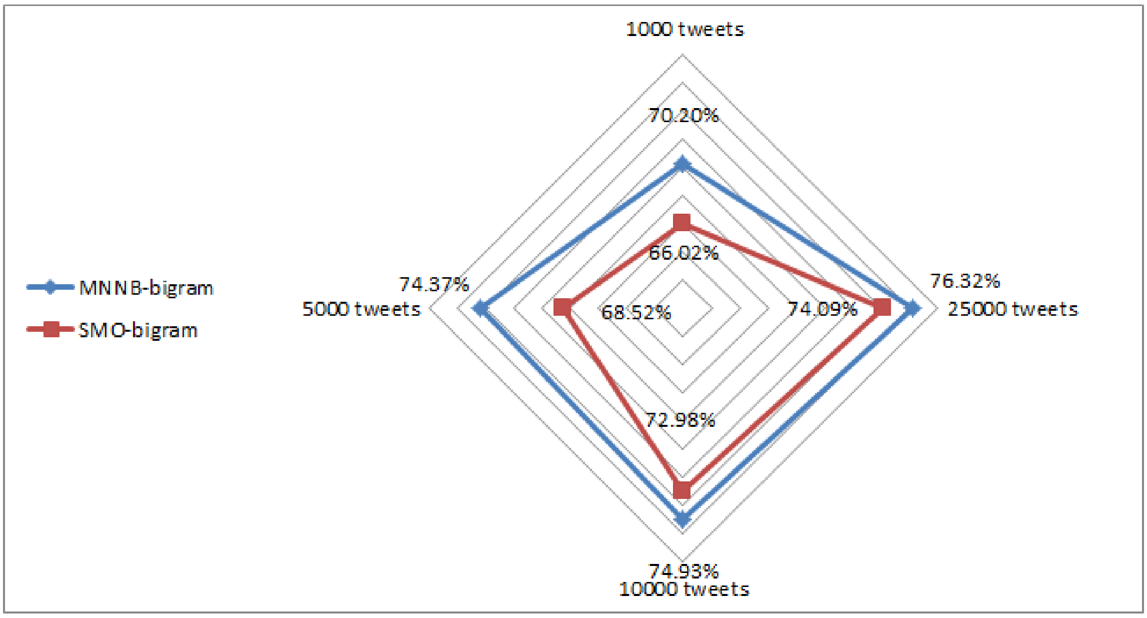


Figure 13 MNNB and SMO classification accuracy using bigram

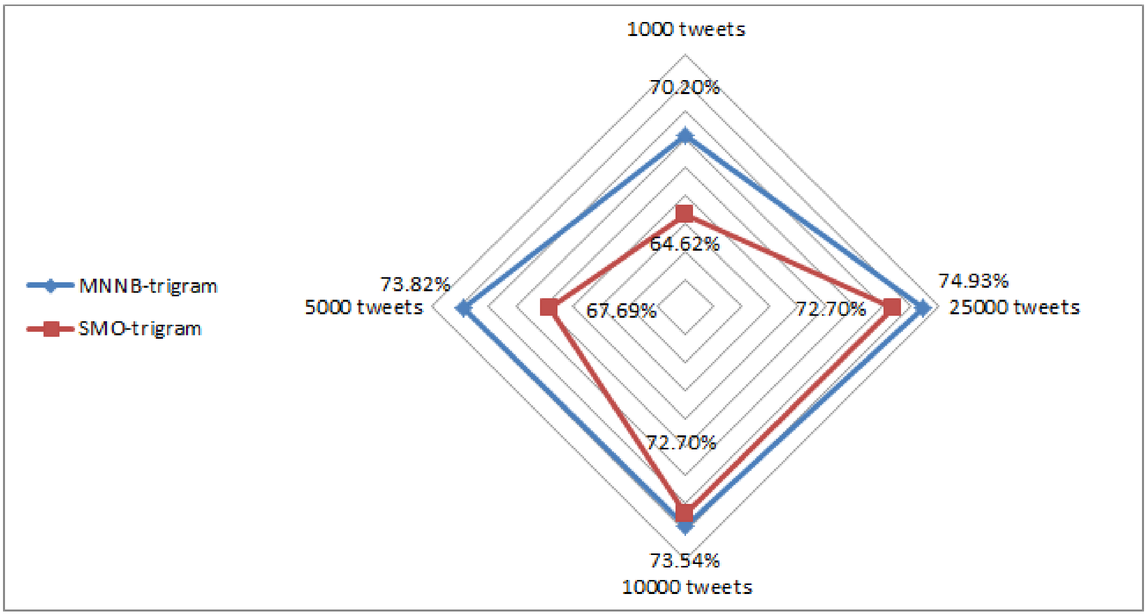


Figure 14 MNNB and SMO classification accuracy using trigram

Therefore after showing the results, it is obvious for the Multi-nominal Naïve-Bayes machine learning model to outperform the Sequential Minimal Optimisation model, approximately through all stages, except through 10,000 tweets in the case of the unigram feature extractor, the SMO outperformed the MNNB with just 1.68 (see Figures 12), Also you can see Figure 15 which represents the accumulative results of the two models. Nevertheless, the accumulative conclusion was showing predomination of the MNNB over SMO not only for the accuracy, but also for the number of tweets that can be treated as one batch per process. Furthermore, the time spent on processing and execution of one big group such as 25,000 tweets may exceed 30 minutes, while for the MNNB, it only lasts not more than 2 minutes.

However, all results obtained from using the two machine learning models (MNNB, SMO) were not very encouraging; however, they answer the questions asked in this study and served the desired objectives of the research.

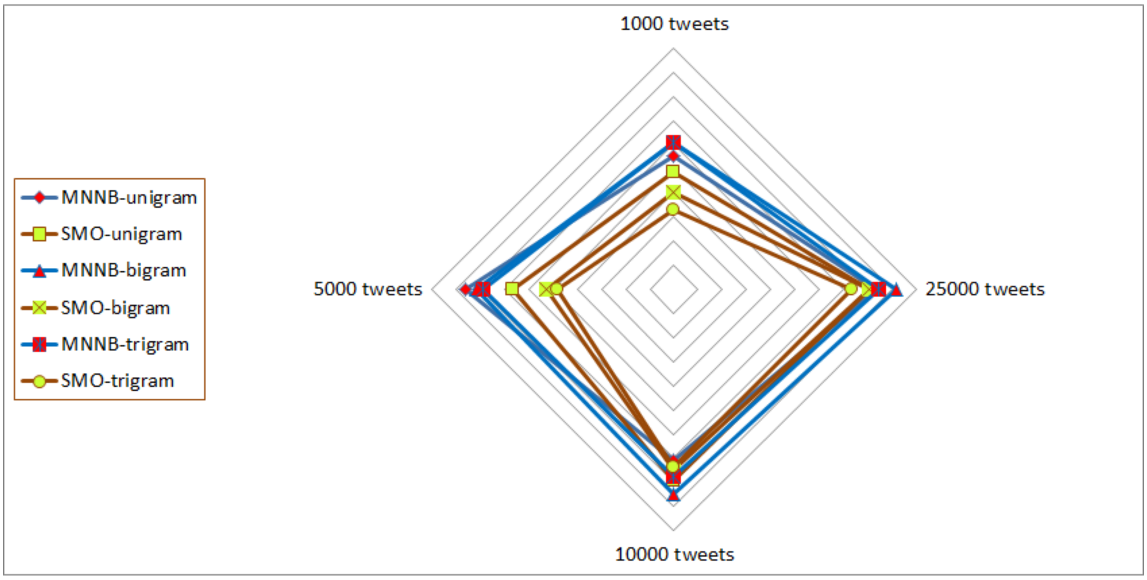


Figure 15 An accumulative scheme obtained from MNNB and SMO.

CHAPTER 5

CONCLUSION

This study has discussed the techniques which have been underlined by most researchers in the field of sentiment analysis and opinion mining. It also explains some important operations which are followed in order to analyse and classify sentiments and emotions, such as opinion and emotion mining operations. The benefits of democratizing data mining through Web 2.0 demonstrate the difference between automated sentiment analysis and human sentiment analysis, and the reason for some companies using a hybrid approach by mixing them together.

In addition to the discussion of the used techniques, this research also offers three findings. The first finding was discussing how to prove the efficiency of using word frequency to classify a document of tweets (multi-domain training dataset) as in the case of using word presence, using unigram MNNB and SMO machine learning models, in the case of multi-domain training corpora.

The second finding here was proving that as long as the number of instances of tweets is increased for the training corpora to classify an external test dataset, one can obtain an accumulative result, high accuracies incrementally using n-gram MNNB and SMO machine learning models.

Finally the research has shown the qualification of using the MNNB model over other models, such as SMO, which is the enhancement of SVM to solve quadric problems that may focus SVM. Moreover, it is easy to apply MNNB on training corpora and test datasets because it deals directly with strings without conversion to other types, whereas in the case of using SMO, it cannot treat the string attribute unless it converts them, for example, to numeric attributes. This will lead to gaining huge feature space and costing more time to process them.

In future works, we hope to fetch more tweets including those from multi-domains so as to increase the classification accuracies obtained from machine learning models. Furthermore, I will make weights for the emoticons rather than remove them, by relying on their position in tweets to provide opportunities for the emoticons to make a contribution to classifying sentiments along with the text of the tweets.

CHAPTER 6

APPENDIX of TABLES

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	no	no	pres.	69.0808%
unigram	removed	no	no	freq.	68.8022%
unigram	removed	no	yes	pres.	67.1309%
unigram	removed	no	yes	freq.	67.1309%

Table 1 (1000 tweets) No stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	yes	no	pres.	67.688%
unigram	removed	yes	no	freq.	69.0808%
unigram	removed	yes	yes	pres.	66.5738%
unigram	removed	yes	yes	freq.	66.5738%

Table 2 (1000 tweets) Stemming was used. (Unigram feature).

Feature	**S.L.**	**Stm.**	**L.C**	**Pres./Freq.**	**MNNB Acc.**
unigram	removed	no	no	pres.	73.2591%
unigram	removed	no	no	freq.	72.9805%
unigram	removed	no	yes	pres.	75.2089%
unigram	removed	no	yes	freq.	74.0947%

Table 3 (5000 tweets) No stemming was used. (Unigram feature).

Feature	**S.L.**	**Stm.**	**L.C**	**Pres./Freq.**	**MNNB Acc.**
unigram	removed	yes	no	pres.	72.7019%
unigram	removed	yes	no	freq.	72.7019%
unigram	removed	yes	yes	pres.	72.4234%
unigram	removed	yes	yes	freq.	71.5877%

Table 4 (5000 tweets) Stemming was used. (Unigram feature).

Feature	**S.L.**	**Stm.**	**L.C**	**Pres./Freq.**	**MNNB Acc.**
unigram	removed	no	no	pres.	72.1448%
unigram	removed	no	no	freq.	71.8863%
unigram	removed	no	yes	pres.	71.5877%
unigram	removed	no	yes	freq.	71.3092%

Table 5 (10000 tweets) No stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	yes	no	pres.	70.7521%
unigram	removed	yes	no	freq.	71.0306%
unigram	removed	yes	yes	pres.	71.0306%
unigram	removed	yes	yes	freq.	70.7521%

Table 6 (10000 tweets) Stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	no	no	pres.	74.6518%
unigram	removed	no	no	freq.	73.5376%
unigram	removed	no	yes	pres.	74.0947%
unigram	removed	no	yes	freq.	73.5376%

Table 7 (25000 tweets) No stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	yes	no	pres.	72.9805%
unigram	removed	yes	no	freq.	72.7019%
unigram	removed	yes	yes	pres.	73.5376%
unigram	removed	yes	yes	freq.	72.4234%

Table 8 (25000 tweets) Stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	no	no	pres.	74.0947%
unigram	removed	no	no	freq.	72.9805%
unigram	removed	no	yes	pres.	76.6017%
unigram	removed	no	yes	freq.	75.766%

Table 9 (50000 tweets) No stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	yes	no	pres.	71.8663%
unigram	removed	yes	no	freq.	72.4234%
unigram	removed	yes	yes	pres.	73.2591%
unigram	removed	yes	yes	freq.	73.8162%

Table 10 (50000 tweets) Stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	no	no	pres.	73.5376%
unigram	removed	no	no	freq.	72.7019%
unigram	removed	no	yes	pres.	76.6017%
unigram	removed	no	yes	freq.	76.8802%

Table 11 (131042 tweets) No stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
unigram	removed	yes	no	pres.	72.4234%
unigram	removed	yes	no	freq.	71.5877%
unigram	removed	yes	yes	pres.	73.5376%
unigram	removed	yes	yes	freq.	72.9805%

Table 12 (131042 tweets) Stemming was used. (Unigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	no	no	pres.	70.195%
bigram	removed	no	no	freq.	69.6379%
bigram	removed	no	yes	pres.	67.4095%
bigram	removed	no	yes	freq.	67.1309%

Table 13 (1000 tweets) No stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	yes	no	pres.	64.9025%
bigram	removed	yes	no	freq.	66.0167%
bigram	removed	yes	yes	pres.	66.0167%
bigram	removed	yes	yes	freq.	65.7382%

Table 14 (1000 tweets) Stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	no	no	pres.	74.3733%
bigram	removed	no	no	freq.	74.3733%
bigram	removed	no	yes	pres.	72.7019%
bigram	removed	no	yes	freq.	72.7019%

Table 15 (5000 tweets) No stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	yes	no	pres.	69.6379%
bigram	removed	yes	no	freq.	70.195%
bigram	removed	yes	yes	pres.	69.3593%
bigram	removed	yes	yes	freq.	70.195%

Table 16 (5000 tweets) Stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	no	no	pres.	74.9304%
bigram	removed	no	no	freq.	74.9304%
bigram	removed	no	yes	pres.	72.7019%
bigram	removed	no	yes	freq.	72.7019%

Table 17 (10000 tweets) No stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	yes	no	pres.	72.1448%
bigram	removed	yes	no	freq.	73.5376%
bigram	removed	yes	yes	pres.	72.9805%
bigram	removed	yes	yes	freq.	73.2591%

Table 18 (10000 tweets) Stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	no	no	pres.	75.766%
bigram	removed	no	no	freq.	74.6518%
bigram	removed	no	yes	pres.	76.3231%
bigram	removed	no	yes	freq.	75.4875%

Table 19 (25000 tweets) No stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	yes	no	pres.	73.2591%
bigram	removed	yes	no	freq.	74.9304%
bigram	removed	yes	yes	pres.	73.2591%
bigram	removed	yes	yes	freq.	72.9805%

Table 20 (25000 tweets) Stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	no	no	pres.	74.6518%
bigram	removed	no	no	freq.	73.5376%
bigram	removed	no	yes	pres.	75.766%
bigram	removed	no	yes	freq.	75.2089%

Table 21 (50000 tweets) No stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	yes	no	pres.	75.2089%
bigram	removed	yes	no	freq.	75.2089%
bigram	removed	yes	yes	pres.	76.3231%
bigram	removed	yes	yes	freq.	76.0446%

Table 22 (50000 tweets) Stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	no	no	pres.	76.8802%
bigram	removed	no	no	freq.	76.8802%
bigram	removed	no	yes	pres.	77.1588%
bigram	removed	no	yes	freq.	77.1588%

Table 23 (131042 tweets) No stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
bigram	removed	yes	no	pres.	75.766%
bigram	removed	yes	no	freq.	76.6017%
bigram	removed	yes	yes	pres.	76.0446%
bigram	removed	yes	yes	freq.	76.6017%

Table 24 (131042 tweets) Stemming was used. (Bigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	70.195%
trigram	removed	no	no	freq.	70.195%
trigram	removed	no	yes	pres.	67.4095%
trigram	removed	no	yes	freq.	67.4095%

Table 25 (1000 tweets) No stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	65.7328%
trigram	removed	no	no	freq.	65.7328%
trigram	removed	no	yes	pres.	66.2953%
trigram	removed	no	yes	freq.	65.7328%

Table 26 (1000 tweets) Stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	73.8162%
trigram	removed	no	no	freq.	73.8162%
trigram	removed	no	yes	pres.	71.8663%
trigram	removed	no	yes	freq.	71.5877%

Table 27 (5000 tweets) No stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	68.2451%
trigram	removed	no	no	freq.	69.6379%
trigram	removed	no	yes	pres.	67.4095%
trigram	removed	no	yes	freq.	68.8022%

Table 28 (5000 tweets) Stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	73.5376%
trigram	removed	no	no	freq.	72.9805%
trigram	removed	no	yes	pres.	71.3092%
trigram	removed	no	yes	freq.	71.5877%

Table 29 (10000 tweets) No stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	71.8663%
trigram	removed	no	no	freq.	72.4234%
trigram	removed	no	yes	pres.	72.9805%
trigram	removed	no	yes	freq.	72.7019%

Table 30 (10000 tweets) Stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	73.8162%
trigram	removed	no	no	freq.	73.2591%
trigram	removed	no	yes	pres.	74.9304%
trigram	removed	no	yes	freq.	74.3733%

Table 31 (25000 tweets) No stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	73.2591%
trigram	removed	no	no	freq.	73.8162%
trigram	removed	no	yes	pres.	73.2591%
trigram	removed	no	yes	freq.	72.9805%

Table 32 (25000 tweets) Stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	73.8162%
trigram	removed	no	no	freq.	73.2591%
trigram	removed	no	yes	pres.	75.766%
trigram	removed	no	yes	freq.	75.766%

Table 33 (50000 tweets) No stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	75.4875%
trigram	removed	no	no	freq.	75.4875%
trigram	removed	no	yes	pres.	75.766%
trigram	removed	no	yes	freq.	76.0446%

Table 34 (50000 tweets) Stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	75.766%
trigram	removed	no	no	freq.	76.0446%
trigram	removed	no	yes	pres.	76.0446%
trigram	removed	no	yes	freq.	76.8802%

Table 35 (131042 tweets) No stemming was used. (Trigram feature).

Feature	S.L.	Stm.	L.C	Pres./Freq.	MNNB Acc.
trigram	removed	no	no	pres.	74.0947%
trigram	removed	no	no	freq.	74.6518%
trigram	removed	no	yes	pres.	74.9304%
trigram	removed	no	yes	freq.	75.2089%

Table 36 (131042 tweets) Stemming was used. (Trigram feature).

Group	Feature	S.L.	Stm.	L.C.	Pres./Freq.	SMO Acc.
1000 tweets	unigram	removed	no	yes	Pres.	67.688%
1000 tweets	unigram	removed	no	yes	Freq.	66.2953%
5000 tweets	unigram	removed	no	yes	Pres.	71.0306%
5000 tweets	unigram	removed	no	yes	Freq.	71.3092%
10000 tweets	unigram	removed	no	yes	Pres.	73.8162%
10000 tweets	unigram	removed	no	yes	Freq.	73.8162%
25000 tweets	unigram	removed	no	yes	Pres.	74.0947%
25000 tweets	unigram	removed	no	yes	Freq.	72.7019%

Table 37 Accuracies of unigram SMO model for all groups

REFERENCES

1. **Wilson T., Wiebe J., Homann P., (2005),** *"Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis",* Published in Proceedings of Human Language Technology and Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, pp. 347-354,

2. **Pang B. and Lee L., (2008),** *"4.1.2 Subjectivity Detection and Opinion Identification",* Opinion Mining and Sentiment Analysis, Published version: Foundations and Trends® in Information Retrieval vol. 2, nos. 1–2, pp. 1–135.

3. **IBM Software Group, Business Analytics, IBM Cognos Consumer Insight, SPSS Predictive Analytics Technology, (2011),** *"http://www.spss.com.ar/MKT/Promos/2012/0512_redessociales/integrate_socialMedia_PA.pdf"*, (Data Download Date: June 03 2014).

4. **Nagarsekar U., Mhapsekar A., Kulkarni P., Kalbande D. R., (2013),** *"Emotion Detection from "The SMS of the Internet""*, IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp. 316 – 321.

5. **Shahheidari S., Dong H., Bin Daud M. N. R., (2013),** *"Twitter Sentiment Mining: A Multi Domain Analysis"*, Published in Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, Taichung, pp.144-149.

6. **Wiebe J., Breck E., Buckley C., Cardie C., Davis P., Fraser B., et al, (2003),** *"Recognizing and Organizing Opinions Expressed in the World Press"*, Paper presented at the Working Notes-New Directions in Question Answering (AAAI Spring Symposium Series), pp. 1–8.

7. **Jindal N., Liu B., (2006),** *"Mining Comparative Sentences and Relations"*, Paper presented at the Proceedings of the 21st National Conference on Artificial Intelligence, (AAAI-06), pp. 1331–1336.

8. **Twitter Sentiment Analysis Using Python and NLTK, (2012),** *"http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/"*, (Data Download Date: January 02 2014).

9. **McCallum A., Nigam K. A, (1998),** *"Comparison of Event Model for Naïve Bayes Text Classification",* In Proceeding of AAAI Workshop on Learning for Text Categorization, Madison (WI US), pp. 41-48.

10. **Pang B., Lee L., Vaithyanathan S., (2002),** *"Thumbs up? Sentiment Classification Using Machine Learning Techniques",* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Stroudsburg, PA, USA, pp.79-86.

11. **Alex W., (2009),** *"Mining the Web for Feelings, Not Facts",* New York Times, 2009-08-23, Retrieved on 2009-10-01, "http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html?_r=2&", (Data Download Date: August 01 2014).

12. **Marshall K., (2009),** *"ReadWriteWeb",* 2009-04-15, Retrieved on 2009-10-01,"http://readwriteweb.com/archives/whats_next_in_social_media_monitoring.php", (Data Download Date October 2 2014).

13. **Cordis, (2010),** *"Collective Emotions in Cyberspace (CYBEREMOTIONS) ",* European Commission, 2009-02-03. Retrieved on 2010-12-13, "http://cordis.europa.eu/project/rcn/89032_en.html", (Data Download Date July 12 2014).

14. **Jamie C., (2010),** *“Flaming Drives Online Social Networks”,* NewScientist, 2010-12-07, Retrieved on 2010-12-13, “http://www.newscientist.com/article/dn19821-flaming-drives-online-social-networks.html#.VMgaUoaSx0g”,(Data Download Date May 26 2014).

15. **Turney P., (2002),** *“Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”,* Published in Proceedings of the 40th Annual Meeting on the Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 417–424.

16. **Snyder B, Barzilay R., (2007),** *“Multiple Aspect Ranking Using the Good Grief Algorithm”,* Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL), pp. 300–307.

17. **Koppel M., Schler J., (2006),** *“The Importance of Neutral Examples for Learning Sentiment”,* presented at Computational Intelligence, vol. 22, issue 2, pp.100-109.

18. **Mike T., Kevan B., Georgios P., Di C., Arvid K., (2010),** *“Sentiment Strength Detection in Short Informal Text”,* Journal of the American Society for Information Science and Technology, vol. 61, issue 12, pp. 2544–2558.

19. **Dey L., Haque S. K. M., (2008),** *“Opinion Mining from Noisy Text Data”,* Published in Proceedings of the second workshop on Analytics for noisy unstructured text data, ACM New York, NY, USA, pp.83-90.

20. **Vu T. T., Pham H. T., Luu C. T., Ha Q. T., (2011),** *“A Feature-Based Opinion Mining Model on Product Reviews in Vietnamese”,* Semantic Methods for Knowledge Management and Communication Studies in Computational Intelligence vol. 381, pp. 23-33.

21. **Zhang L., Liu B., (2014),** *“Aspect and Entity Extraction for Opinion Mining”,* Data Mining and Knowledge Discovery for Big Data Studies in Big Data Vol. 1, pp. 1-40.

22. **Jindal N., Liu B., (2008),** *"Opinion Spam and Analysis",* In Proceeding of 1'ACM International Conference on the Web Search and Data Mining, Stanford CA, ACM, New York, NY, pp. 219-229.

23. **Liu B., (2012),** *"Sentiment Analysis and Opinion Mining",* Ch. 8, electronic Google book, , pp. 113-117.

24. **Bollen J., Mao H., Zeng X. J., (2011),** *"Twitter Mood Predicts the Stock Market",* Journal of Computational Science, vol. 2, no. 1, pp. 1-8 .

25. **Khan K., Baharudin B., Khan A., (2009),** *"Mining Opinion from Text Documents: A Survey",* in Digital Ecosystems and Technologies, DEST'09, 3rd IEEE International Conference on, Istanbul, pp. 217–222.

26. **Mishne G., (2005),** *"Experiments with Mood Classification in Blog Posts",* In Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access, Citeseer, pp. 19.

27. **Padmaja S., Prof. Fatima S. S., (2013),** *"Opinion Mining and Sentiment Analysis: An Assessment of Peoples' Belief: A Survey",* International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) vol.4, no.1, pp. 21-33.

28. **Su F., Markert K., (2008),** *"From Words to Senses: a Case Study in Subjectivity Recognition",* In Proceedings of Coling, Manchester, UK, *"*http://www.comp. leeds.ac.uk/markert/Papers/Coling2008.pdf*"*, pp. 1-8.

29. **Hu M., Liu B., (2004),** *"Mining and Summarizing Customer Reviews",* Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168–177.

30. **Liu B., (2010),** *"Sentiment Analysis and Subjectivity",* Handbook of Natural Language Processing, pp. 627-666.

31. **Kim S.M., Hovy E.H., (2006),** *"Identifying and Analysing Judgment Opinions",* Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006), New York, NY, pp. 1-8.

32. **Erin W., (2013),** *"Human Sentiment Analysis",* Growing Social Media. Retrieved 14 November 2013, "http://growingsocialmedia.com/human-sentiment-analysis/", (Data Download Date: September 16 2014).

33. ***Ogneva M., (2010),*** *is the Director of Social Media at Biz360, "How Companies Can Use Sentiment Analysis to Improve Their Business",* http://mashable.com/2010/04/19/sentiment-analysis/, (Data Download Date: September 18 2014).

34. **Strapparava C., Valitutti A., (2004),** *"WordNet-Affect: An Affective Extension of WordNet",* In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, pp. 1083-1086.

35. **Baccianella S., Esuli A., Sebastiani F., (2010),** *"Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining",* In Proceedings of LREC, pp. 2200–2204.

36. **Cambria E., Havasi C., Hussain A., (2012),** *"SenticNet 2: A Semantic and Affective Resource for Opinion mining and Sentiment Analysis",* In Proceedings of AAAI FLAIRS. pp. 202–207.

37. **Case Study, (2013),** *"Advanced Sentiment Analysis",* "http://paragonpoll.com/sentiment-analysis-systems-case-study/", pp. 1-3.

38. **Galitsky B., McKenna E. W., (2013),** *"Sentiment Extraction from Consumer Reviews for Providing Product Recommendations",* "http://docs.google.com/viewer?url=patentimages.storage.googleapis.com/pdfs/US20090282019.pdf", (Data Download Date: July 07 2014).

39. **Nute D., (1994),** *"Defeasible logic",* In Handbook of logic in artificial intelligence and logic programming, volume 3: Non-monotonic reasoning and uncertain reasoning, Oxford University Press, pp. 353-395.

40. **Antoniou G., Billington D., Governatori G., Maher M. (2001),** *"Representation results for defeasible logic",* ACM Transactions on Computational Logic, vol.2, issue 2, pp. 255-287.

41. **Galitsky B., Dobrocsi G., de la Rosa J. L., (2010),** *"Inverting Semantic Structure under Open Domain Opinion Mining",* In Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference FLAIRS Conference, pp. 186-191.

42. **Galitsky B., Chen H., Du S., (2009),** *"Inversion of Forum Content Based on Authors' Sentiments on Product Usability",* AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0, pp.33–38.

43. **Ogneva, M., (2012),** *"How Companies Can Use Sentiment Analysis to Improve Their Business",* "http://mashable.com/2010/04/19/sentiment-analysis/", (Data Download Date: October 09 2014).

44. **Ekman P., (1992),** *"An Argument for Basic Emotions",* Cognition & Emotion, 6(3-4), pp. 169-200.

45. **Ma C., Prendinger H., and Ishizuka M., (2005),** *"Emotion Estimation and Reasoning Based on Affective Textual Interaction",* Affective computing and intelligent interaction, Beijing, China, pp.622-628.

46. **Chuamartin, F., (2007),** *"A knowledge-Based System for Headline Sentiment Tagging",* In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval), Association for Computational Linguistics, , Prague, pp. 422-425.

47. **Andreevskaia A., Bergler S., (2006),** *"Mining Word-Net for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses",* In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), Trento, IT, pp. 209-216.

48. **Cai L., Hofmann T., (2003),** *"Text Categorization by Boosting Automatically Extracted Concepts",* Paper presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ACM New York, NY, USA, pp. 182-189.

49. **Alm C.O., Roth D., Sproat R., (2005),** *"Emotions from Text: Machine Learning for Text-Based Emotion Prediction",* In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 579-586.

50. **Wilson T., Wiebe J., Hwa R., (2004),** *"Just How Mad are You? Finding Strong and Weak Opinion Clauses",* Published In AAA '04 Proceedings of 19th National Conference on Artificial Intelligence, 2004, pp. 761-769.

51. **Yu B., Kaufmann S., Diermeier D., (2008),** *"Exploring the Characteristics of Opinion Expression for Political Opinion Classification",* In Proceedings of the international conference on Digital government research, Digital Government Society of North America, pp. 82-91.

52. **Melville P., Gryc W., Lawrence R. D., (2009),** *"Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification",* In Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, pp. 1275-1284.

53. **Daye K., Lawrence S., Pennock D.M., (2003),** *"Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews",* In Proceedings of the 12th international conference on World Wide Web, ACM, pp. 519-528.

54. **Hatzivassiloglou V., McKeown K.R., (1997),** *"Predicting the Semantic Orientation of Adjectives",* In Proceedings of the eight conference on European chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 174-181.

55. **Donkor B., (2013),** *"Social Media Analytics", "*http://brnrd.me/social-sentiment-sentiment-analysis*",* (Data Download Date: August 07 2014).

56. **Guadette L., Japkowicz N., (2011),** *"Compact Features for Sentiment Analysis",* Published In Canadian AI'11 Proceeding of the 24th Canadian conference on Advances In Artificial Intelligence, Springer-Verlag Berlin, Heidelberg, pp. 146-157.

57. **Yousefpour A., Ibrahim[1] R., Hamd[2] H. N. A., (2014),** *"A Novel Feature Reduction Method in Sentiment Analysis",* International Journal of Innovative Computing, vol. 4, no. 1, pp. 34-40.

58. **Twitter Help Centre, (2014),** *"https://support.twitter.com/entries/166337-the-twitter-glossary".* (Data Download Date: August 09 2014).

59. **Lovins J. B., (1968),** *"Development of a Stemming Algorithm",* Electrical System Laboratory, Massachusetts Institute of Technology, Massachusetts 02139, Mechanical Translation and Computational Linguistics, vol.11, nos. 1and 2, pp. 22-31.

60. **Domingos P., (2012),** *"A Few Useful Things to Know about Machine Learning",* University of Washington, Seattle, published in Magazine Communications of the ACM , vol. 55 issue 10, ACM New York, NY, USA, pp. 78-87.

61. **Huan W., (2007),** *"Exploring Intrinsic Structures from Samples: Supervised, Unsupervised and Semisupervised Frameworks",* Master of Philosophy in Information Engineering, Chapter 1, The Chinese University of Hong Kong, pp. 1-3.

62. **Sutton R. S., Barto A. G., (1998),** *"Reinforcement Learning: An Introduction",* a Bradford Book, the MIT Press, Cambridge, Massachusetts, London-England, Ch. 1, pp.5-7.

63. **Kibriya A. M., Frank E., Pfahringer B., Homles G., (2005),** *“Multinominal Naïve Bayes for Text Categorization Revisited”,* Department of Computer Science, Waikato University, Hamilton, New Zealand, Published at Springer, Berlin, pp. 488-499.

64. **Platt J., (1998),** *“Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines”,* CiteSeerX: 10.1.1.43.4376., pp. 1-21.

APPENDICES A

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Abd, Mustafa

Date and Place of Birth: 16 November 1980, Baghdad

Marital Status: Married

Phone: +964 7901 432211

Email: mus.sal.abd@gmail.com

EDUCATION

Degree	**Institution**	**Year of Graduation**
M.Sc.	Çankaya University, Information Technology-Ankara.	2015
Higher Diploma	Technology University, Computer Science, Artificial Intelligence Department-Baghdad.	2003
B.Sc.	Baghdad University, College of Science, Computer Science Department-Baghdad.	2002

WORK EXPERIENCE

Year	Place	Enrollment
2012 Present	Baghdad University, College of Science, Computer Science Department-Baghdad.	Lecturer
2011-2012	Ministry of Higher Education and Scientific Research, Department of Scholarships and Cultural Relations, Section of Evaluation and the Equivalence of Diplomas	Calculating Rates
2009-2011	Baghdad University, College of Science, Computer Science Department-Baghdad.	Lecturer
2007-2009	Baghdad University, Akhawarzmi Engineering College, The division of Planning, Studies, and follow-up.	Manager assistant
2005-2006	Baghdad University, Psychological Research Center	Web Designer - Maintenance

FOREIN LANGUAGES

Good English, Beginner Turkish.

PROJECTS

Web site of Architectural domes: circular and conical in Iraq-Ministry of Tourism. 2002.

HOBBIES

Listen Classic Music, Travel, Walking, Read Poetry, Meditation.