CLASSIFICATION OF NEGATION IN SENTIMENT ANALYSIS USING TWITTER DATA

A thesis submitted in partial fulfillment of the requirement for the

The degree of Master of Computer Application

Of

Jadavpur University

By

ANIK DIAN

Registration Number: 133672 of 2015-2016

Examination Roll Number: MCA186010

Under the Guidance of

Dr. Diganta Saha

Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

May 2018

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled “CLASSIFICATION OF NEGATION IN SENTIMENT ANALYSIS USING TWITTER DATA” has been satisfactorily completed by Anik Dian (University Registration No.: 133672 of 2015-16, Examination Roll No.:MCA186010).It is a bonafide piece of work carried out under my guidance and supervision and be accepted in partial fulfillment of the requirement for the Degree of Master of Computer Application, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, Kolkata.

Dr. Diganta Saha (Thesis Supervisor)

Professor
Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Countersigned

Dr. Ujjwal Maulik

Professor
Head, Department of Computer Science and Engineering,
Jadavpur University, Kolkata-700032.

Prof. Chiranjib Bhattacharjee
Dean, Faculty of Engineering and Technology,

Jadavpur University, Kolkata-700032.

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled “CLASSIFICATION OF NEGATION IN SENTIMENT ANALYSIS USING TWITTER DATA” is a bonafide record of work carried out by Anik Dian in partial fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University during the period of February 2018 to May 2018. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, the opinion expressed or conclusion drawn there in but approve the thesis only for the purpose for which it has been submitted.

____

Signature of Examiner

Date:

____

Signature of Supervisor

Date:

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis entitled “CLASSIFICATION OF NEGATION IN SENTIMENT ANALYSIS USING TWITTER DATA” contains literature survey and original research work by the undersigned candidate, as part of his Degree of Master of Computer Application.

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 materials and results that are not original to this work.

Name: Anik Dian
University Registration No. : 133672 of 2015-16
Examination Roll No. : MCA186010

Thesis Title: CLASSIFICATION OF NEGATION IN SENTIMENT ANALYSIS USING TWITTER DATA

Signature

Date:

ACKNOWLEDGEMENT

First and foremost, I would like to start by thanking God Almighty for showering me with the strength, knowledge and potential to embark on this wonderful journey and to persevere and complete the embodied research work satisfactorily.

I am pleased to express my deepest gratitude to my thesis guide, **Dr. Diganta Saha**, **Professor**, Department of Computer Science and Engineering, Jadavpur University, Kolkata for his invaluable guidance, constant encouragement and inspiration during the period of my dissertation.

I am highly indebted to **Jadavpur University** for providing me the opportunity and the required infrastructure to carry on my thesis.

I am thankful for all the teaching and non-teaching staff whose helping hands have smoothed my journey through the period of my research.

Last but not the least; I would like to thank my family members, classmates, seniors and friends for giving me constant encouragement and mental support throughout my work.

Anik Dian
University Registration No. : 133672 of 2015-16
Examination Roll No. : MCA186010
Master of Computer Application
Department of Computer Science and Engineering
Jadavpur University

Table of Contents

Abstract

In the past decade, new forms of communication, such as microblogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them. In this work, system discussed a paradigm to extract the sentiment from a famous microblogging service, Twitter.Tweets are unambiguous short texts messages that are up to a maximum of 140 characters, where users post their opinions for everything and the analysis of twitter dataset with data mining approach such as use of Sentiment Analysis Algorithm using Lexicon Based Algorithms and Machine Learning Algorithms. I have proposed a system that analyzed tweets about three categories- positive, negative and neutral. For this I have used API based approach. Then I have used Dictionary-Based Approach and Supervised Learning. This proposed system classifies the negations in sentiment analysis using Twitter data and 81.11 percentage of accuracy has found.

Keywords: *sentiment, negation, polarity, emoticons.*

Chapter One

Introduction

1.1 What is Sentiment?

Generally the term ‘sentiment’ defines people's opinions or feelings about a situation or a general feeling, attitude, or opinion about something or athought, opinion, or idea based on a feeling about a situation, or a way of thinking about something.
Generally, a binary opposition in opinions is assumed as for/against, like/dislike, good/bad, etc. So ‘Sentiment’ means feelings, attitudes, emotions, Opinions, subjective impressions. Using NLP, statistics, or machine learning methods to extract, identify or otherwise characterize the sentiment content of a text unit sometimes referred to as opinion mining, although the emphasis, in this case, is on extraction.

1.2 What is Sentiment Analysis?

Opinion mining (sometimes known as Sentiment Analysis or emotion AI) refers to the use of Natural Language Processing (NLP), text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study effective states and subjective information [16]. Sentiment Analysis is widely applied to the voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Generally speaking, Sentiment Analysis aims to determine the attitude of a speaker, writer, or another subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation, affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor).

A basic task in Sentiment Analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy"[33].

1.3 Application of Sentiment Analysis

There are many application of Sentiment Analysis like:

1. Adjust marketing strategy

Most companies, if not all, are active in social media, and use the public forum to promote their brands and services. But how can you know if you are doing the right things in social media? From the managerial perspective**,** social media is not just a platform where you can post and promote your services. It is a place where your customer's chit chats about your brand and is full of information about how brand is being perceived by your target customers. The information you get from Sentiment Analysis provides you with means to optimize your marketing strategy. By listening to what your customers feel and think about your brand, you can adjust your high-level messaging to meet their needs. From the tactical point of view, you can build a short-term marketing campaign to provide customers with what they want. By continuously having sentiment analysis in place, you can adjust your campaign to fit even more to your target audience [37].

2. Measure ROI of your marketing campaign

Success of your marketing campaign is not measured only by the increase in the number of followers, likes, or comments. The success also lies in how many positive discussions you are able to help facilitate amongst your customers. By doing Sentiment Analysis, you can to see how much positive or negative discussions have occurred amongst your audience. By combining the quantitative and the qualitative measurements, you can measure the true ROI of your marketing campaign.

3. Develop product quality

Sentiment Analysis helps you complete your market research by getting to know what your customers' opinions are about your products/services and how you can align your products/services' quality and features with their tastes. Your products and services are judged not only by how well it performs functionally but also by how nicely it is presented in the forms of, for example, beautiful package design, irresistible promotions, reasonable pricing, and even impressive store decoration. Ideas to develop your product quality and how it is presented can only be derived from your target customers' opinions. One way to do that is by conducting a structured and planned survey. Another method is to get that information from the casual discussions that are going on related to your brand in public social platforms.

4. Improve customer service

Once your customer purchases your product, you want to keep them loyal to your brand as long as possible, and be an evangelist for your brand. Your customers can essentially become micro-influencers for you. That is why it is of utmost importance to have the best customer service in place and keep your current customers happy.

There are many factors that contribute to great customer service, such as on-time delivery, being responsive in social media, and adequate compensation for product's errors. Sentiment Analysis can pick up negative discussions, and give you real-time alerts so that you can respond quickly.

5. Crisis management

Constant monitoring of what is currently happening in social media conversations also helps you to prevent or at least mitigate the damage of online communication crisis. Crisis might stem from your product's bad quality, unacceptable customer service, or other serious social issues such as environmental harm, animal cruelty or child labor usage in emerging markets. If you don’t manage customer complaints fast enough, the conversation can go viral and lead to a huge crisis that you might not be able to cope back from. Uber is a great example of recent crisis that went viral in social media [37].

7. Sales Revenue

The biggest benefit of doing Sentiment Analysis (and even doing business in the first place) is to boost sales revenue. I listed this benefit as the last point in this blog because the increase in sales revenue is the final outcome of successful marketing campaigns, improved products/service quality, and customer service, which can be achieved with sentiment analysis.

1.4 Objective of Sentiment Analysis

The objective of Sentiment Analysis is to find out the positive, negative or neutral feelings, emotions and opinions written in a text. These sentiments are based on the meaning of words used in text according to different scenarios and situations. There are a variety of ways used to express the same feeling in a written text by using different grammatical rules. These grammatical rules contain negations that are very frequently used in text that completely change the meanings of words. In other words, negation identification, detecting and classifying its scope within a sentence (text) are necessary in finding out the sentiments from a piece of text. Although classification of negation is an important aspect of Sentiment Analysis [24], it is yet to be properly addressed. In general, the efforts put into Sentiment Analysis of sentences having negation terms in them are less efficient with respect to general Sentiment Analysis. Classification of negation is not a simple task and its complexity increases, since negation words such as not, nor etc., (syntactic negation) are not the only criterion for negation calculation. The linguistic patterns - prefixes (e.g., un-, dis-, etc.) or suffixes (e.g., -less) also introduce the context of negation in textual data. Similarly, word intensifiers and diminishes (contextual valence shifter) also flip the polarity of sentiments. It will take a lot of efforts to enlist all such words in one list. These valence shifters do not only flip the polarity but also increase negativity. On the other hand, negation does not restrict itself to ‘not’. There are terms like; no, not, n’t, never, no longer, no more, no way, nowhere, by no means, at no time, etc that also change the meaning of a sentence. Another reason is the fact that the number of negation sentences encountered is considered insignificant during the evaluation of any Sentiment Analysis system as compared to the level of effort required to resolve the issues related to negation [15]. So, in this work I have classified negations using Naïve Bayes and calculated accuracy.

1.5 Importance of Sentiment Analysis

Social media sentiment analysis can be an excellent source of information and can provide insights that can:

- Determine marketing strategy
- Improve campaign success
- Improve product messaging
- Improve customer service
- Test business KPIs
- Generate leads

The study of Sentiment Analysis, if done properly, is exceptionally complex and is actually a field of study, not just a feature in a social media tool. I should probably be clear at this point, that the objective of this blog is not to discuss the nuances and detail of Sentiment Analysis. In fact, quite the opposite, I am trying to simplify this very complex topic so that you can use the information when deciding on the social media tool or services that you need.

1.6 Sentiment analysis Terminology

In this segment, we need the various terms used in the Sentiment Analysis.

Fact: A fact is that which has truly happened or which is really the case.

Opinion: An opinion is a view or judgment formed about something (like Product or movie) not necessarily based on fact or knowledge.

Subjective Sentence: A sentence or a text is a subjective or opinionated if it actually indicates one's feelings or emotions.

Objective Sentence: An objective sentence indicates some facts and known Information about the world. For example: universal truths.

Review: A review is texts that contain a particular combination of words that have opinions of customer a particular item or opinions of viewers for a movie. A review may be subjective or objective or even both.

Item: An individual article or unit, especially one that is part of a list, collection or set.

Known Aspects: Known aspects are default aspects provided by the certain website for which users separately give ratings.

Sentiment: Sentiment is a polarity term that implies to the direction in which a behavior or opinion is expressed. For example, excellent is a sentiment for the attribute camera in the sentence "the camera of the phone is excellent".

Opinion Polarity: Opinion Polarity or Subjectivity Orientation denotes the polarity expressed by the user or customer or viewer in terms of numerical values. Rating: Most of the people use star ratings for expressing polarity, represented by stars in the range from 5 to 1 which is called ratings.

Polarity: Polarity is a three-way orientation scale. In this, a sentiment can be either negative or positive or neutral.

1.7 What is Negation and types of Negation?

The presence of the word negation is able to change the polarity of the text if it is not handled properly it will affect the performance of the sentiment classification. Negation is a complex phenomenon that has been studied from many perspectives, including cognition, philosophy, and linguistics. As described by Lawler (2010, page 554) [18], cognitively, negation “involves some comparison between a ‘real’ situation lacking some particular element and an ‘imaginable’ situation that does not lack it.” In the logic formalisms, “negation is the only significant monadic factor,” whose behavior is described by the Law of Contradiction that asserts that no proposition can be both true and not true? In natural language, negation functions as an operator, like quantifiers and models [15]. A main characteristic of operators is that they have a scope, which means that their meaning affects other elements in the text. The affected elements can be located in the same clause (5a) or in a previous clause (5b).

(5) a. We didn’t find the book.
(5) b. We thought we would find the book. This was not the case.

The study of negation in philosophy started with Aristotle, but nowadays is still a topic that generates a considerable number of publications in the field of philosophy, logic, psycholinguistics, and linguistics. “An analysis of negation in relation to semantic and pragmatic phenomena”, Tottie (1991) studies negation as a grammatical category from a descriptive and quantitative point of view, based on the analysis of empirical material [1]. She defines two main types of negation in natural language: rejections of suggestions and denials of assertions. Denials can be explicit and implicit. Languages have devices for negating entire propositions (clausal negation) or constituents of clauses (constituent negation). Most languages have several grammatical devices to express clausal negation, which are used for different purposes like negating existence, negating facts, or negating different aspects, modes, or speech acts (Payne 1997). As described by Payne (page 282): …. a negative clause is one that asserts that some event, situation, or state of affairs does not hold. Negative clauses usually occur in the context of some presupposition, functioning to negate or counter-assert that presupposition [5]. Van der Wouden (1997) defines what a negative context is, showing that negation can be expressed by a variety of grammatical categories [6]. We reproduce some of his examples in Example (6).

(6) a. Verbs: We want to avoid doing any look-up, if possible.
(6) b. Nouns: The positive degree is expressed by the absence of any phonic sequence.
(6) c. Adjectives: It is pointless to teach any of the vernacular languages as a subject in schools.
(6) d. Adverbs: I've never come across anyone quite as brainwashed as your student.
(6) e. Prepositions: You can exchange without any problem.
(6) f. Determiners: This fact has no direct implications for any of the two methods of font representation.
(6) g. Pronouns: Nobody walks anywhere in Tucson.
(6) h. Complementizers: Leave the door ajar, lest any latecomers should find themselves shut out.
(6) i. Conjunctions: But neither this article nor any other similar review I have seen then had the methodological discipline to take the opposite point of view.

Negation can also be expressed by affixes, as in motionless or unhappy, and by changing the intonation or facial expression, and it can occur in a variety of Syntactic Constructions. Typical negation problems that persist in the study of negation are determining the scope when negation occurs with quantifiers (7a), neg-raising (7b), the use of polarity items (7c) (any, the faintest idea), double or multiple negation (7d), and affixed negation (Tottie, 1991) [1].

(7) a. All the boys didn't leave.
(7) b. I don't think he is coming.
(7) c. I didn't see anything.
(7) d. I don't know anything no more.

Negations can be classified as [20]:
Syntactic-no, not, rather, couldn't, wasn't, didn't, wouldn't, shouldn't, weren't, don't, doesn't, haven't, hasn't, won't, wont, hadn't, never, none, nobody, nothing, neither, nor, nowhere, isn't, can't, cannot, mustn't, mightn't, shan't, without, needn't
Diminishes -hardly, less, little, rarely, scarcely, seldom
Morphological Prefixes: de-, dis-, il-, im-, in-, ir-, mis-, non-, un- , Suffix: -less

1.8 What is Polarity and types of Polarity?

Polarity, also known as orientation is the emotion expressed in the sentence. It can be positive, negative or neutral. It simply means emotions expressed in a sentence.Emotions are closely related to sentiments. The strength of a sentiment or opinion is typically linked to the intensity of certain emotions, e.g., joy and anger.Opinions in sentiment analysis are mostly evaluations (although not always).

According to consumer behavior research, evaluations can be broadly categorized into two types [38]:

1. Rational Evaluation

2. Emotional Evaluation

Rational evaluation: Such evaluations are from rational reasoning, tangible beliefs,and utilitarian attitudes. For example, the following sentences

For example "The voice of this phone is clear," "This car is worth the price," and "I am happy with this car."

Emotional evaluation: Such evaluations are from non-tangible and emotional responses to entities which go deep into people's state of mind.

For example, the following sentences express emotional evaluations: "I love iPhone," "I am so angry with their service people" and "This is the best car ever built."

To make use of these two types of evaluations in practice, it can be designed 3 sentiment ratings, negative (-1), neutral (0) and positive (+1). In practice, neutral often means no opinion or sentiment expressed.

1.9 What isSubjectivity?

Subjectivity is when text is an explanatory article which must be analyzed in context. The subjectivity of sentence may express some personal feelings, views, or beliefs. It can be also defined by Linguistic Expression of somebody's opinions, sentiments, emotions, beliefs, evaluations, speculations [33].Examples of Subjective Expressions:

References to private states:"She was enthusiastic about the plan."

References to speech or writing events expressing private states:"Leaders rounding condemned his verbal assault on Israel."

Expressive subjective elements: "What a freak show."

1.10 What areEmoticons and types of Emoticons?

Emoticons are ASCII art, sometimes these are called “Smileys or Emojies “. These emoticons (e.g., :) and :() have been widely used in sentiment analysis and other NLP tasks as features of Machine Learning Algorithms or as entries of sentiment Lexicons. In this work, I argue that while emoticons are strong and common signals of sentiment expression on social media the relationship between emoticons and sentiment polarity are not always clear. Thus, an algorithm that deals with sentiment polarity should take emoticons into account but extreme caution should be exercised in which emoticons to depend on. For researchers and businesses, having access to its huge amount of user-generated data is critical for understanding user behavior and the sentiment expressed.Emoticons, such as :) ;) :-) and :(, are frequentlyused online in social media, IM (e.g., Skype), blogs,forums, and other kinds of online social interactions. Becausethey are commonly used in online communicationsand they are often direct signals of sentiment [32]. Different online communities and tools may elicit varying degrees of emoticon usage. Emoticons can categorize into three types: Positive emoticons, Negative emoticons, Neutral emoticons. I have got some texts containing positive emotions, such as happiness, amusement or joy, some containing negative emotions, such as sadness, anger or disappointment or some containing neutral emoticons. In Figure 1, I give some emoticons and their meaning [40].

Emoticons	Meaning
:)	Happy Face
:D	Laugh
;)	Wink
:(	Sad Face
:|	Straight Face
:@	Angry
DX	Screaming

Figure 1: Some Emoticons and their meaning

Emoticons are changing the way we communicate faster than linguists can keep up with or lexicographers can regulate. Today Emoticons are used widely in social media, some reasons are given below:

1. Users react by Emoticons like they would real human face.
2. Emoticons are even okay in Business settings.
3. Emoticons soften the blow of a critique.
4. Emoticons make you appear more friendly and competent.
5. Emoticons co-related with real life happenings.

Chapter Two

Literature Reviews

2.1 Introduction

A literature review discusses state of the art in a particular subject area. It can be an analysis of literature or published sources, on a particular topic.A literature review is a simple summary of the sources, but it usually has an organizational pattern and combines both summary and synthesis. A summary is a recap of the important information of the source, but a synthesis is a re-organization, or a reshuffling, of that information. It might give a new interpretation of old material or combine new with old interpretations. Or it might trace the intellectual progression of the field, including major debates. And depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant.

2.2 Survey of Related Work

In recent years a lot of work has been done in the field of “Sentiment Analysis on Twitter “by a number of researchers. In its early stage, it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative only. They proposed either machine learning approach or lexicon-based approach or they may even include a combination of both to achieve good accuracy. The classification of polarity items along semantic lines, in terms of their licensing environments, has been a topic of debate for the last 30 years or so.

❑ The first proposal on Negations was introduced by Frans Zwarts in 1981, in a paper in Dutch, called “Negatief Polaire Uitdrukkingen”. In that paper, he proposed a binary distinction between weak and strong items, the weak items being licit in all downward entailing contexts, the strong ones in a proper subset thereof, the set of contexts determined by what he called anti-additive functions. Zwarts’ paper was inspired by the dissertation of Bill Ladusaw, which appeared a year before (Ladusaw 1979). Ladusaw had introduced the notion of downward entailment as the key to generalizing over all contexts in which polarity items like any and every shows up [2].

❑ In the year 1993 Talmy Givón proposed a function based English grammar [3]. It covers the grammatical subsystems commonly found in simple clauses: Verbal inflections, auxiliaries and the grammar of tense-aspect modality and negation; articles determiners, pronouns and the grammar of referential coherence; the variety of noun phrases and noun modifiers.

❑ Wendy et al proposed a simple algorithm for identifying negated findings and diseases in discharge summaries in the year 2001[7].

❑ Turney et al (2002) used bag-of-words method for sentiment analysis in which the relationships between words were not at all considered and a document is represented as just a collection of words. To determine the sentiment for the whole document, sentiments of every word was determined and those values are united with some aggregation functions [8].

❑ Pang Lee et al (2002) classified documents not by topics but by sentiments, e.g. determining whether the review is positive or negative. For negation handling, if a word x follows the negation word then a new feature 'NOT x' created tag every word from x until first punctuation mark. But this method cannot model the scope of negation, because it is heuristically tagging all word until it finds the mark, without concerning with negation words or not. Addition in preprocessing task, mostly the punctuation marks are removed; this is for simplification in preprocessing stage [9].

❑ Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task (Turney, 2002[8]; Pang and Lee, 2004[9]), it has been handled at the sentence level (Hu and Liu, 2004[11]; Kim and Hovy, 2004) and more recently at the phrase level (Agarwal et al., 2009) [13].

❑ Parikh and Movassate (2009) [14] implemented two models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model.

❑Jia et al. (2009) [15] studied the impact of each occurrence of a negation term in a sentence on its polarity and introduced the concept of scope of the negation terms.

❑ Pak and Paroubek (2010) [16] proposed a model to classify the tweets as positive and negative. They created a Twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, they developed a sentiment classifier based on the multinomial Naive Bayes method that uses features like N-gram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

❑ Bifet and Frank (2010) [17] used Twitter streaming data provided by Firehose API, which gave all messages from every user which are publicly available in real-time. They experimented multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They arrived at a conclusion that SGD-based model, when used with an appropriate learning rate was the better than the rest used.

❑ Davidov et al.,(2010) [18] proposed an approach to utilize Twitter user-defined hashtags in tweets as a classification of sentiment type using punctuation, single words, n-grams and patterns as different feature types, which are then combined into a single feature vector for sentiment classification. They made use of K-Nearest Neighbor strategy to assign sentiment labels by constructing a feature vector for each example in the training and test set.

❑ Xia et al., (2011) [21] used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In their work, they used two types of feature sets (Part-of-speech information and Word-relations) and three base classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines). They applied ensemble approaches like the fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy.

❑ In 2012, Balakrishnan Gokulakrishn et al. [22] proposed an approach where a plugged stream of tweets from the Twitter microblogging webpage are preprocessed and grouped in light of their emotional content as positive, negative and irrelevant; and investigate the execution of different ordering calculations in light of their precision and recall in such cases.

❑ Calvinet. al., (2014) [26] proposed a model where sentiment extremity of Twitter surveys are measured utilizing Naïve Bayes classifier strategy. The model demonstrates a promising come about on characterizing the ubiquity in light of consumer satisfaction and along these lines characterizing the best supplier to be utilized.

❑ Aizhan Bizhanovaet. al., (2014) [27] proposed a model for naturally characterizing the opinion of Twitter messages toward item/mark, utilizing emoticons and by enhancing preprocessing steps keeping in mind the end goal to accomplish high exactness.

❑ Po-Wei Liang et al. (2014) [28] used Twitter API to collect twitter data. Their training data falls into three different categories (camera, movie, mobile). The data is labeled as positive, negative and non-opinions. Tweets containing opinions were filtered. Unigram Naive Bayes model was implemented and the Naive Bayes simplifying independence assumption was employed. They also eliminated useless features by using the Mutual Information and Chi-square feature extraction method. Finally, the orientation of a tweet is predicted i.e. Positive or Negative.

❑ Pablo et al. [29] presented variations of Naive Bayes classifiers for detecting the polarity of English tweets. Two different variants of Naive Bayes classifiers were built namely Baseline (trained to classify tweets as positive, negative and neutral), and Binary (makes use of a polarity lexicon and classifies as positive and negative. Neutral tweets neglected). The features considered by classifiers were Lemmas (nouns, verbs, adjectives,and adverbs), Polarity Lexicons, and Multiword from different sources and Valence Shifters.

❑ As research in Indonesia, Bojar (2015) [30] who conducted research about the resources of the lexicon for Indonesian sentiment also did the negation handling. By adopting the technique from Das and Chen [8] handled the negation of sentiment caused by a negation word. Bojar uses negation words such as 'tidak', 'tak', 'tanpa', 'belum', and 'kurang'. The words that occur between the negation words and the first punctuation after the negation word are tagged with 'NOT_'. Example, there is a sentence: 'kameranya kurang bagusgam barnya' became 'kameranya kurang NOT_bagus NOT_gambarnya'.

❑ For the negation in sentiment, there are some of the researchers that focus on the impact of the negation in sentiment sentences. A survey conducted by Wang et al (2015) [32], they survey for negation role in sentiment analysis. They state that effective negation model for sentiment analysis usually requires the knowledge of polar expression.

❑ Sanjana Woonna et al.(2016) [33] proposed a framework that examinations tweets into three classifications which are positive, negative and neutral utilizing supervised learning approach After the execution, the outcomes demonstrated which viewpoints individuals like or aversion and how feelings on motion pictures change over a timeframe.

❑ In 2016, Sandip D Mali et al. [34] proposed another framework called SentiView which a vocabulary-based approach for sentiment investigation. They have gotten high accuracy because of preprocessing and expulsion of non-opinion tweets from data.

❑ In 2017, Kai Yang et Al [36] proposed a highly effective hybrid model combining different single models to overcome their weaknesses. They build the sentimental dictionary from exterior data. As single model have many limitations and weakness. That way they build a hybrid model by combing many single approaches to overcome those limitations of the single model. The experimental results show that our hybrid model shows very great performance. In hybrid model 2 approaches that are SVM and GDBT (Gradient boosting decision tree) are combined together that is based on stacking approach.

2.3Conclusion

Therefore, I decided to work on Classification of Negation in Sentiment Analysis Using Twitter data. In this work, I used Naïve Bayes Classifier and calculate the accuracyof classification of Negation using this algorithm.

Chapter Three

Sentiment Analysis Techniques

3.1 Academic Prospective of Sentimental Analysis

Sentiment Analysis (SA) or Opinion Mining (OM) – is a discipline that has seen a lot of activity since about 2000. [1] The verge expansion and associability of the social media's and proliferation of social media and its tools (e.g. Twitter, Facebook, LinkedIn, etc.), that has made the accessibility to information about how people feel about things more readily available to the masses. The main fields of research in Sentimental Analysis are Subjectivity Detection, Sentiment Prediction,Aspect-Based Sentiment Summarization, Textual Summarization, Constructive Viewpoint Summarization, Opinion based-entity ranking, Review Detection, Product Feature Extraction, OpinionRetrieval.

- Subjectivity Detection is about determining if a piece of text actually contains opinions or not (i.e. subjective expression or objective?)[33].

- Sentiment Prediction is specifically about predicting the polarity as it is positive or negative or neutral at the literal level.

- Aspect-Based Sentiment Summarization is to provide a sentiment summary of service or product at the feature or aspect level (i.e. start rating or service score).

- Textual Summarization gathers a few informative sentences or phrases that summaries the review of the product.

- Constrictive Viewpoint Summarization is to highlight the contradiction in opinions.

- Opinion based-entity ranking is the task of ranking entities based on opinions [24].

- Review Detection deals with identifying the real comments and about to identifying the fake opinion from reviews.

- Product Feature Extraction is a task to extract the product features from its review.

- Opinion Retrieval is a task of searching a specific opinion and gathers it.

3.2 Industrial Prospective of Sentimental Analysis

Sentiment Analysis techniques are classified into two categories namely Lexicon Based Approach and Machine Learning Based Approach.

Lexicon Based Approach is further divided into two categories namely Dictionary Based and Corpus-Based approach. In Dictionary-Based Approach, the sentiment is identified using synonym and antonym from lexical dictionary like WordNet. In Corpus-Based Approach, it identifies opinion words by considering word list. Corpus-BasedApproach furthermore classified as a statistical and semantic approach. In statistical approach, co-occurrences of words are calculated to identify sentiment. In semantic approach, terms are represented in semantic space to discover the relation between terms [34].

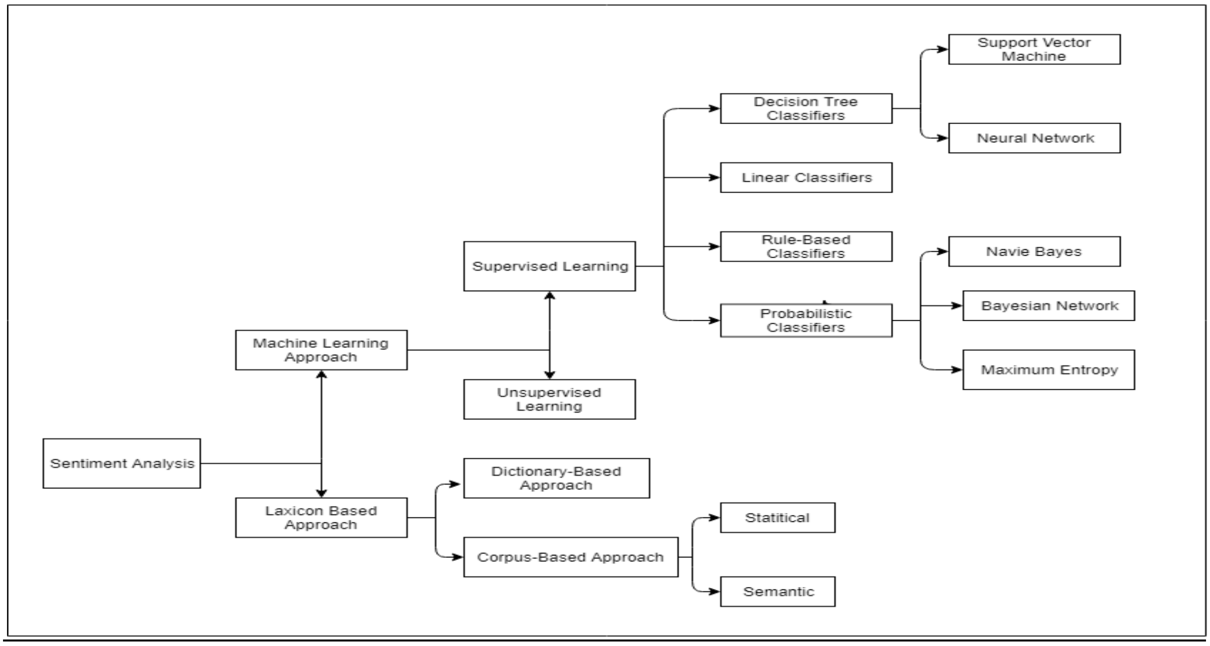


Figure 2. Sentiment Analysis techniques [33]

3.2.1 Machine Learning Approach

Machine Learning is a field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed. Machine Learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. The text classification methods using Machine learning are divided into Supervised and Unsupervised learning methods.

- The Supervised Learning Methods use a large number of the training dataset.
- The Unsupervised Learning Methods are used when it’s difficult to find in training dataset.

3.2.1.1 Supervised Learning

The Supervised Learning depends on the existence of previously labeled dataset. In next sub-section, I present brief details of some classifiers, used in the analysis.

3.2.1.1.1 Probabilistic Classifiers

Probabilistic Classifiers are considered to be among the most popular classifiers for the Machine Learning. This is developed by assuming generative models which are product distributions over the original attribute space (as in Naive Bayes) [10]. It uses mixture model for classification. It is also called generative classifiers because a generative model is a model for generating all values for a phenomenon. The used two classifiers are discussed in next sections.

3.2.1.1.1.1 Naive Bayes Classifier (NB)

Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions. It is based on “Bayes Theorem”. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some per define the finite set. In machine learning, naive Bayes classifiers are a family of simple probabilistic Classifiers based on applying Bayes' theorem with strong (Naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of Parameters linear in the number of variables (features/predictors) in a learning problem [10].

Maximum likelihood training can be done by evaluating a closed form expression (mathematical Expression that can be evaluated in a finite number of operations), which takes linear time.

It is based on the application of the Bayes rule given by the following formula:

$$P(C{=}c|\ D{=}d) = \frac{P(D=d|C=c)P(C=c)}{P(D=d)}$$

And the simplified Bayes formula can be written as:

$$P\ (c|\ d) = \frac{P(d|c)P(c)}{P(d)}$$

c -Hypothesis, d -Tuples, P (c|d) represents Posterior probability of c conditioned on d i.e. the Probability that a Hypothesis holds true given the value of d, P(c) represents Prior probability of c i.e. the Probability that c holds true irrespective of the tuple values, P(d|c) represents posterior Probability of d conditioned on c i.e. the Probability that d will have certain values for a given Hypothesis, P(d) represents Prior probability [10].

In our case, a tweet d is represented by a vector of K attributes such as d = (w, w... w) Computing P(d|c) is not trivial and that's why the Naive Bayes introduces the assumption that all of the feature values wj are independent given the category label c. That is, for i =/j, wi and wj are conditionally independent given the category label c.So the Bayes rule can be rewritten as:

$$P(c|d) = \frac{P(c) \times \prod_{j=1}^{k} p(w(j)|c)}{P(d)}$$

3.2.1.1.1.2 Maximum Entropy Classifiers

The Max Entropy classifier is a probabilistic classifier to the class of exponential models. The Max Entropy does not assume that the features are conditionally independent of each other. It works on principle of Maximum Entropy and its output has the Maximum Entropy.The Max Entropy Classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more. Maximum Entropy maximizes the entropy that is defined in the conditional probability distribution where c is the class, d is the

Tweet. The formula is given below:

$$P_{ME}(\mathrm{C|D}) = \frac{exp[\sum \lambda_i f_i(c,d)]}{\sum exp[\sum \lambda_i f_i(c,d)]}$$

The Maximum Entropy Classifier (known as a conditional exponential classifier) converts labeled feature set to vector using encoding. This encoded vector is then used to calculate weights for each feature that can be then be combined to determine the most likely label for a feature-set by an f {encoding}.In particular, the encoding maps each f{feature set, label} pair to a vector[23].

3.2.1.1.2 Linear Classifiers

Given $\overline{X}$= $\{x_1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x_n\}$ is the normalized document word frequency, vector $\overline{A}$= $\{a_1 \ldots\ldots\ldots\ldots\ldots a_n\}$ is a vector of linear co-efficients with the same dimensionality as the fiture space , and b is a scalar; the output of the linear predictor is define as p = $\overline{XA} + b$, which isthe output of the Linear Classifiers; among them is Support Vector Machine (SVM)[4] which is a form of classifiers that attempt to determine good linear separators between different classes.

3.2.1.1.2.1 Neural Network (NN)

Neural Network is inspired by biological neural networks in the brain. The neural network Systems can learn to do tasks by being trained on examples, rather than task-specific programming Comprised of layers of interconnected nodes that get activated by an activation function and connections are dealt with by the propagation function. Multiple Neural Networks are used for non-linear boundaries. Multilayer of Neural Network is used to induce multiple piecewise linear boundaries. In Neural Network, the feed of a layer neuron is the output of its previous network.The Neural Network training is a complex process. The Neural Network can be used for text classification and perception learning [26].

3.2.1.1.2.2 Support Vector Machine (SVM)

In Machine Learning, Support Vector Machines (SVM, also support vector networks) are Supervised Learning models with associated learning algorithms that analyze data used for classification and regression analysis. Support Vector Machine (SVM) developed by Vladimir Vapnik and it was first heard in 1992, introduced by Vapnik, Boser and Guyon in COLT-92 [4].SVM became popular because of its success in handwritten digit recognition in NIST (1998).

3.2.1.2 Weakly, Semi and Unsupervised Learning

The main purpose of text classification is to classify a document into a certain number of pre-defined categories. To accomplish this, a large no of labeled data (training set) are used in Supervised Learning.Unsupervised Machine Learning is the Machine Learning task of inferring a function to describe hidden structure from "unlabeled" data [39].

- Weak supervision is supervision with noisy labels. For example, bootstrapping, where the bootstrapping procedure may mislabel some examples.
- Semi-supervised learning is when you have a dataset that is partially labeled and partially unlabeled.
- Unsupervised learning is a type of Machine Learning Algorithm used to draw inferences from datasets consisting of input data without labeled responses.

In Unsupervised Methods, the data set is split into small packets (keys) and categorized each sentence using keyword list of each category and sentence similarity measure.

3.2.2 Lexicon Based Approach

Semantic Orientation (SO) is a measure of subjectivity and opinion in text. It usuallycaptures an evaluative factor (positive or negative) and potency or strength (degreeto which the word, phrase, sentence, or document in question is positive or negative).towards a subject topic, person, or idea.Application of a lexicon is one of the two main approaches to sentiment analysis and it involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text. With this approach, a dictionary of positive and negative words is required, with a positive or negative sentiment value assigned to each of the words. The semantic orientation of phrases is determined as positive if it is more related to "best" and is considered too negative if it is more it's related to "poor". It is based on opinion lexicon [16].

- The Dictionary-Based Approach which depends on finding option seed words and search a dictionary of their synonyms and antonyms.
- The Corpus-Based Approach starts with a seed list of opinion words and then finds other opinion words in a large corpus to help in finding opinion words with context-specific orientations.

3.2.2.1 Dictionary-Based Approach

In Dictionary-Based Approach, a small set of opinion words are collected. It is English database dictionary where every term is associated with each other via the link. Mostly WordNet is used to check similarity with words and to calculate sentiment score. It links to sets of the syntactic category which are verb, adjective, adverb,and noun. WordNet and Dictionary based approach; both are improved and add new entries (newly found word) after completion of each iteration. It is linked with semantic relations those are termed as a synonym, antonym, hyponymy, metonymy, troponymy, Entailment etc [4].

The major disadvantage of Dictionary-Based Approach is the inability to fine opinion words in domain and context specific orientation. It is used for texting advertisement and it improves irrelativeness of user experience.

3.2.2.2 Corpus-Based Approach

Corpus linguistics is the study of language as expressed in corpora (samples) of "real world" text. The text-corpus method is a digestive approach that derives a set of abstract rules that govern a natural language from texts in that language and explores how that language relates to other languages. Originally derived manually, corpora now are automatically derived from source texts. Corpus linguistics proposes that reliable language analysis is more feasible with corpora collected in the field in its natural context ("realia"), and with minimal experimental-interference. Corpus linguistics has generated a number of research methods, which attempt to trace a path from data to theory [24].

3.3 Conclusion

In this work, I have used Machine Learning Approach and Lexicon Based Approach. In Machine Learning Approach I have used Probabilistic Classifier: Naïve Classifier and in Lexicon Based Approach I have used Dictionary-Based Approach.

Chapter Four

Proposed Method

4.1 Introduction

There are two main techniques for classification of negation in sentiment analysis using Twitter data: Machine Learning Based and Lexicon Based. Firstly I stood up with the API Based Approach. This approach has done to analyze the sentiment of the tweets. Twitter data were collected and given as input in API system. It has given output as positive, negative or neutral sentiment. After analyzing the sentiments hide in the text I have done Dictionary-Based Approach and cluster the words and emoticons. After completing Dictionary based approach I have done Machine Learning Approach. There are two types of Machine Learning Approach: Unsupervised and Supervised. I have done a Probabilistic Classification (using Naive Bayes Classifier) under Supervised Approach.

4.2 Proposed Method

The proposed system is divided into following steps:

4.2.1 Twitter Data Collection

Firstly I have collected a large no of the Twitter dataset (100000 tweets) by using Twitter API to do this Sentiment Analysis process. Twitter bases its application programming interface (API) off the Representational State Transfer (REST) architecture. REST architecture refers to a collection of network design principles that define resources and ways to address and access data [41]. The architecture is a design philosophy, it's not only set of blueprints, there's no single prescribed arrangement of computers, servers, and cables. For Twitter, a REST architecture in part means that the service works with most Web syndication formats. So basically, Twitter API is simply a set of URLs that take parameters. They URLs let you access many features of Twitter, such as posting a tweet or finding tweets that contain a word, etc. The first way to access Twitter data is through their public API. Making calls using the Twitter API is similar to using Twitter's search feature but allows you to get this data in an automated fashion. If you need access to data beyond what's possible with Twitter's search and real-time APIs, Twitter also offers Firehose. This API is near real-time, guarantees 100% uptime and provides direct access to data about individual users and profiles. This incredible data stream is only available to a limited number of distributors [41].

In this process, I have used AYLINE text Analysis API. I have created a Twitter Developer account and registered my application. During this process I have received a consumer key and a consumer secret: these are used in application settings and from the configuration page of application we also require an access token and an access token secret which provide the application access to Twitter on behalf of the account. The process is divided into two sub-processes [41]. These are discussed in next subsection.

4.2.1.1 Accessing Twitter Data

I have made my own application by using Twitter API and to interact with Twitter services I have used Twitter provided REST API. I have used a bunch of Python-based clients.The API variable is now my entry point for most of the operations which can be performed with Twitter. The API provides features to access different types of data. In this way, we can easily collect tweets (and more) and store them in the system. By default, the data is in JSON format, I have changed it to text format for easy accessibility.

4.2.1.2 Streaming

Power Track, Volume (e.g. Decahose, Firehose), and Replay streams utilize Streaming HTTP protocol to deliver data through an open, streaming API connection. Rather than delivering data in batches through repeated requests by your client app, as might be expected from a REST API, a single connection is opened between your app and the API, with new results being sent through that connection whenever new matches occur. This results in a low-latency delivery mechanism that can support very high throughput.After establishing the connection to stream, I have begun receiving a stream of data. The body of the response consists of a series of carriage-return (\r\n) delimited JSON-encoded activities, system messages, and blank lines. By extending and customize the stream-listener process, I have processed the incoming data from JSON to Text format. This way, I gathered more than 100000 tweets in the month of February (From 18thFeb to 22th Feb). This is especially true for live events with a worldwide live coverage.

4.2.2 Data Pre-Processing and Cleaning

One of the first steps in working with text data is to pre-process it. It is an essential step before the data is ready for analysis. Majority of available text data is highly unstructured and noisy in nature – to achieve better insights or to build better algorithms, it is necessary to play with clean data. This Twitter data is highly unstructured, it is an informal communication – typos, bad grammar, usage of slang, the presence of unwanted content like URLs, Stopwords, and Expressions etc. are the usual suspects. In this step, therefore I discuss these possible noise elements and how I cleaned them step by step. I am providing ways to clean data using Python. There are some key attributes presents in the previously collected dataset.

- text: the text of the tweet itself
- created_at: the date of creation
- favorite_count, retweet_count: the number of favorites and retweets
- favorite, retweeted: Boolean stating whether the authenticated user (you) have favorite or retweeted this tweet
- Lang: acronym for the language (e.g. “en” for English)
- id: the tweet identifier
- place, coordinates, geo: geo-location information if available
- user: the author’s full profile
- entities: list of entities like URLs, @-mentions, hashtags and symbols
- in_reply_to_user_id: user identifier if the tweet is a reply to a specific user
- in_reply_to_status_id: status identifier id the tweet is a reply to a specific status

I have applied an extensive set of pre-processing steps to decrease the size of the feature set to make it suitable for learning algorithms. The cleaning method is based on Dictionary methods. Beside this, some more steps are required in this process [13]. These are described below:

4.2.2.1 HTML Character Escaping

Data obtained from twitter usually contains a lot of HTML entities like &lt; &gt; &amp; which gets embedded in the original data. It is thus necessary to get rid of these entities. One approach is to directly remove them by the use of specific regular expressions. Hare, we are using the HTML parser module of Python which can convert these entities to standard HTML tags. For example&lt; is converted to “<” and &amp; is converted to “&”. After this, we remove this special HTML Character and links.

4.2.2.2 Decoding Data

This is the process of transforming information from complex symbols to simple and easier to understand characters. The collected data uses different forms of decoding like “Latin”, “UTF8” etc. I have changed all of this to “UTF-8” for better understanding.

4.2.2.3 Stop Word Removal

Stop words are generally thought to be a “single set of words”. I should not want these words taking up space in our database. I have used NLTK and a “Stop Word Dictionary” to remove the stop words as they are not useful.

4.2.2.4 Removal of Punctuations

All the punctuation marks according to the priorities should be dealt with. For example: “.”, “,”,”?” are important punctuations that should be retained while others need to be removed. I have replaced every word boundary by a list of relevant punctuations present at that point. I have also removed any single quotes that might exist in the text.

4.2.2.5 Other Words

In the twitter datasets, there is also other information as retweet, Hashtag, Username and Modified tweets. All of this is ignored and removed from the dataset.

4.2.2.6 Junk Tweets

The junk tweets were detected and deleted as they are useless in further processing using API. The “junk” label means that the tweet cannot be understood by a human annotator and many of these tweets were not translated well using Google translate. After removal of junk tweets, I got around 90000 tweets.

4.2.2.7 Tokenize Twitter Text

Tokenization is the act of breaking up a sequence of strings into pieces such as meaningful words, keywords, phrases, symbols and other elements called tokens. The list of tokens is became input for the fartherprocess.In tokenization, each sentence is split up into small parts.The tokenization is based on regular expressions (regexp), which is a common choice for this type of problem. Some particular types of tokens (e.g. phone numbers or chemical names) will not be captured and will be probably broken into several tokens. To overcome this problem, as well as to improve the richness of your pre-processing pipeline, you can improve the regular expressions, or even employ more sophisticated techniques like Named Entity Recognition.I have used “nltk.tokenize” and import “word_tokenize” to tokenize twitter data.

4.2.3 Counting

After completing cleaning and Tokenization I have collected 2000 tweets in the English language randomly from the previously collected dataset. I have done a process of counting the total number of words, number of positive words, number of negative words, number of negative contractions and number of the acronym in that twitter dataset.I have written a simple python program for counting. The input of this python script was 2000 tweets and data dictionary (Positive word dictionary, negative word dictionary, negative contraction dictionary and acronym dictionary) and the output is given below:

Total number of tweets collected	2000
Total number of words	28935
Total number of positive words	365
Total number of negative words	809
Total number of negative contractions	287
Total number of acronym	632

Figure 3 – Counting of words

From the above result, this is clear that the percentage of negations (negative contractions and negative words) used in twitter data set is 3.78% which is much higher, whereasa percentage of positive words used in Twitter data is 1.26%. From this statistical survey, it is obvious that negations have more impact in sentiment analysis.

I have done another counting process which is based on Emoticons. I have used emoticons dictionary as an input of a simple python script. The script computes with the dataset of 90000 tweets and finds the total number of Emoticons. I have found 7458 emoticons in this dataset.Figure 4 shows the most frequent emoticons and their frequency in the data set. “:)” is most frequently (46.4%) used emoticons. As expected, many of the emoticons were used infrequently. I have selected the emoticons that occurred more than 0.8%, which results in 24 emoticons given in the next table:

Emoticons	Counting	Emoticons	Counting
:)	3460(46.4%)	‘:]	111(1.5%)
:(	1066(14.3%)	-_-	100(1.46%)
:D	857(11.5%)	=-s	84(1.13%)
;)	462(6.2%)	;(	82(1.1%)
:-=	305(4.1%)	:,’[	74(1.0%)
:@	290(3.9%)	:((	74(1.0%)
xD	238(3.2%)	:-/	70(0.97%)
:P	216(2.9%)	=]	67(0.9%)
(:	137(1.84%)	<3=>	67(0.9%)
:|	134(1.8%)	:”D	60(0.82%)
:>	126(1.7%)	=D	60(0.82%)
]:D	124(1.69%)	DX	59(0.8%)

Figure 4: Counting of emoticons

4.2.4 Sentiment Analysis

At first, I have analyzed sentiments in the tweets through API and then calculateda number of positive, negative and neutral tweets in the data set of 2000 tweets.

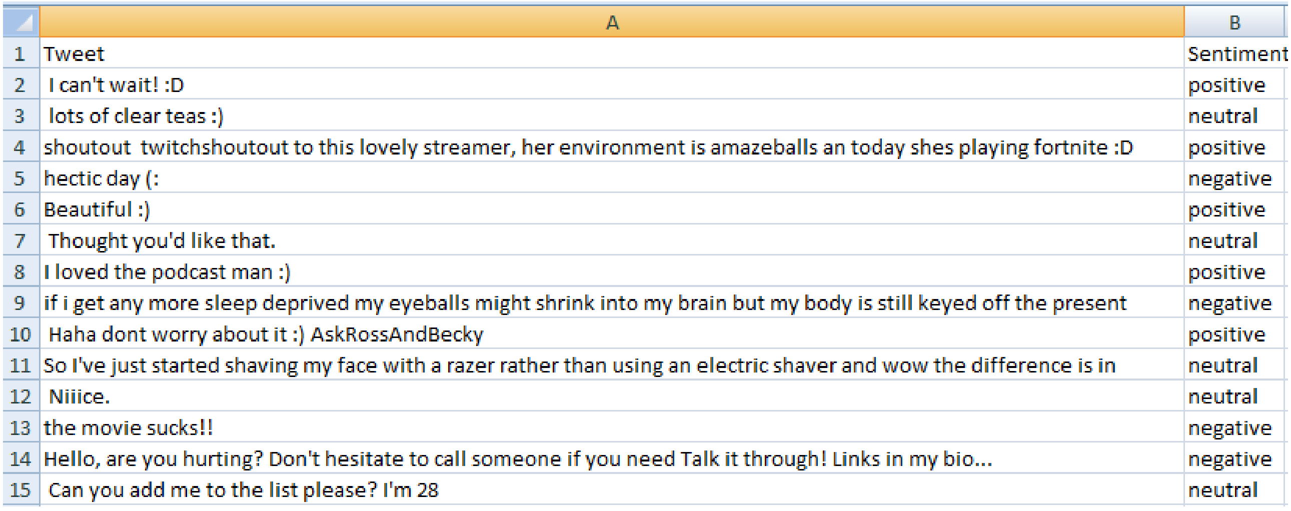
4.2.4.1 API Based

I have used AYLINE text analysis API. This text API consists of eight distinct Natural Language Processing, Information Retrieval,and Machine Learning APIs which when combined, allow developers to extract meaning and insight from any document with ease. This API does the following process to extract sentiment from a Twitter [41].

- Article Extraction
- Article Summarization
- Classification using IPTC NewsCode Standards
- Entity Extraction
- Concept Extraction
- Language detection
- Sentiment Analysis

4.2.4.2 Result and analysis

The results of the analysis are shown through the screen-shot given below:



	A	B
1	Tweet	Sentiment
2	I can't wait! :D	positive
3	lots of clear teas :)	neutral
4	shoutout twitchshoutout to this lovely streamer, her environment is amazeballs an today shes playing fortnite :D	positive
5	hectic day (:	negative
6	Beautiful :)	positive
7	Thought you'd like that.	neutral
8	I loved the podcast man :)	positive
9	if i get any more sleep deprived my eyeballs might shrink into my brain but my body is still keyed off the present	negative
10	Haha dont worry about it :) AskRossAndBecky	positive
11	So I've just started shaving my face with a razer rather than using an electric shaver and wow the difference is in	neutral
12	Niiice.	neutral
13	the movie sucks!!	negative
14	Hello, are you hurting? Don't hesitate to call someone if you need Talk it through! Links in my bio...	negative
15	Can you add me to the list please? I'm 28	neutral

Figure 5: screen-shot of sentiment analysis through API

I have counted a number of positive, negative and neutral tweets from the above analysis.

Total Number of tweets	2000
Number of positive tweets	223
Number of negative tweets	595
Number of neutral tweets	1182

Figure 6: Statistical analysis

It is seen that the percentage of negative tweets is 33.13% i.e. almost one-third of the whole twitter data.

4.2.5 Clustering of Words and Emoticons

A cluster is a group of objects that belong to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. A cluster of data objects can be treated as one group.While doing cluster analysis, first I have partitioned the set of data into groups based on data similarity and then assign the labels to the groups.The main advantage of clustering over classification is that it is adaptable to changes and helps single out useful features that distinguish different groups.I used “Lexicon-Based Approach” to do the clustering.

4.2.5.1 Lexicon-Based Approach

4.2.5.1.1 Introduction

A Lexicon Based Approach is a simple, viable and practical approach to Sentiment Analysis of Twitter data without training. In this approach, I have used previously collected data set and analyzed it by Dictionary-Based Approach. I have to feed the system four types of the dictionary for doing the clustering, dictionaries are given as:

- An Acronym dictionary of more than 500 acronyms with their translation.
- A Positive and Negative word dictionary of more than 10000 words with given polarity (sentiment-out-of-context) of each word.
- Negative contractions and auxiliary's dictionary of which will be detected negation in a given tweet such as don't, can't etc.
- An Emoticon dictionary with more than 1000 emoticons.

4.2.5.1.2 Result and analysis

I have done clustering of words taken from the dataset of 2000 tweets. I have taken the help from previously collected dictionaries and WordNet (version 2.1) to cluster the words. I have collected all the positive words, negative words, negative contractions and acronym manually by consulting with dictionary and WordNet (to know the sense of the word). The resultant clusters of the words are given below:

A	achievements, adore, amazed, awesome, awesome, beautiful, best, best-performing, charming, dazzling, darling, enjoyment, fav, gorgeous, good, happy, magnificent, NC, satisfactory, sweetheart, win, yummy
B	advantage, bliss, calm, cheerful, cool, enjoy, friendly, kindly, quiet, ready, sporty
C	bff, lol, a little bit, less positive,gn, not easy, not bad, unbiased, bad dream, move-on, not satisfactory
D	btw,b4,cre8,da,home,hello, selfiii, smart phone , stuck , smart phone ,u, What
E	aren't, can't, couldn't, haven't, hadn't, shouldn't, wasn't, won't
F	angry, awful, bad, bitch, bloody, break-up, cry, chaotic, damn, deadly, fuck, kills, suck, scandal, stupid
G	bullshit,careless,cheat,disappointments,lie,mistakes,odd,painfull,sack,voulnarable

Figure 7: clustering of words

Cluster A, B, C contains mostly positive words, cluster D contains mostly neutral words, E, F contains mostly negative words and cluster G contains mostly slangs. During this analysis, I have got some common spelling mistakes done by users like lose (correct-loose), alot (correct-a lot), confusion between affect and effect, to and too. I have found that use of acronym words are in large numbers. Uses of negative contractions are also very high. In Figure 5, I give some popular negative contractions and their full form:

Negative contractions	Meaning
can't	can not
couldn't	could not
hadn't	had not
wasn't	was not
Won’t	will not

Figure 8: Some negative contractions

I have found various types of negations used by users. Some of them are given below:

- Auxiliary Negation:

There is a negation rule in English: If we want to state that something is not true, we can form a negative sentence by adding the word “not” after the first auxiliary verb in the affirmative sentence.

Example: I can’t study r8 now.

- Noun Phrase Negation:

Another way of changing an affirmative sentence into a negative sentence is to place a negative determiner or a restrictive quantifier before a noun, which is called noun phrase negation.

Example: Not manypeople came to the meeting last week.

- Adverb Negation:

In English, there are some negative adverbs which create negative sentences, without adding any negative expression. So, when using a negative adverb, we don’t need the ‘no’ part of a negative sentence.

Example: She never apologizes for her wrong behavior.

- Morphological Negation**:**

This type of negation is also called affixal negation and is marked by the presence of negative affixes: a-, non-, dis-, un-, in- (including the variants im-, il-, it-), the suffix-less and the suffix -out.

Example: useless fellow.

I have done the clustering of emoticons with the help of Emoticons Dictionary. The result is given below:

A	:D :) xD
B	;) =D :”D :> : P <3=>
C	(: =]
D	:| -_-
E	:(:((:@ :,’[
F	]:D :-= ‘:]
G	=-S DX

Figure 9: clustering of emoticons

Figure 9, shows that the emoticons in the clusters A, B, C most positive emoticon and cluster D include neutral emoticon group whereas E, F, G include a mostly negative emoticon. The percentage of negative emoticons (evaluated from Figure 3) is 26.79% which is quite high. The emoticon helps us to understand the words in the same cluster in the terms of the sentiment they expressed. In the summary, this clustering result helps us to validate that the emoticons were often used inconsistent context to help express sentiment and sentiment express by emoticons agrees well with words around them [1].

4.2.5.2 Machine Learning Based Approach

4.2.5.2.1 Introduction

Machine Learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.It allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output value within an acceptable range.I have used Waikato Environment for Knowledge Analysis (WEKA). WEKA is a machine learning tool. The input of this WEKA tool is a CSV file with 1096 negations and their subjectivity, polarity,and cluster (E, F and G). For model building, we applied supervised machine-learning algorithms: Naïve Bayes (Probabilistic Classifier) on the training dataset. This Machine Learning Algorithms (Naïve Bayes) is applied to the training set to build an analysis model. On the basis of the model constructed for the analyzer, the test set was evaluated. After test set evaluation, I have recorded the accuracy of the analyzer under this model.

4.2.5.2.2 Naïve Bayes Classifier (With Training Dataset)

I have considered 1096 negations and calculated the polarity and subjectivity of these negations using a TextBlob. I have made the training data set through the following steps.

STEP 1: Create CSV file

1. Create an EXCEL file with all 1096 negations.
2. Calculate Subjectivity and Polarity of each negation using TextBlob and plot in the above EXCEL file labeled with respective negation.
3. Create a CSV file from the above EXCEL file where each negation labeled with their Polarity, Subjectivity and the cluster (E, F,and G) (Figure 5).

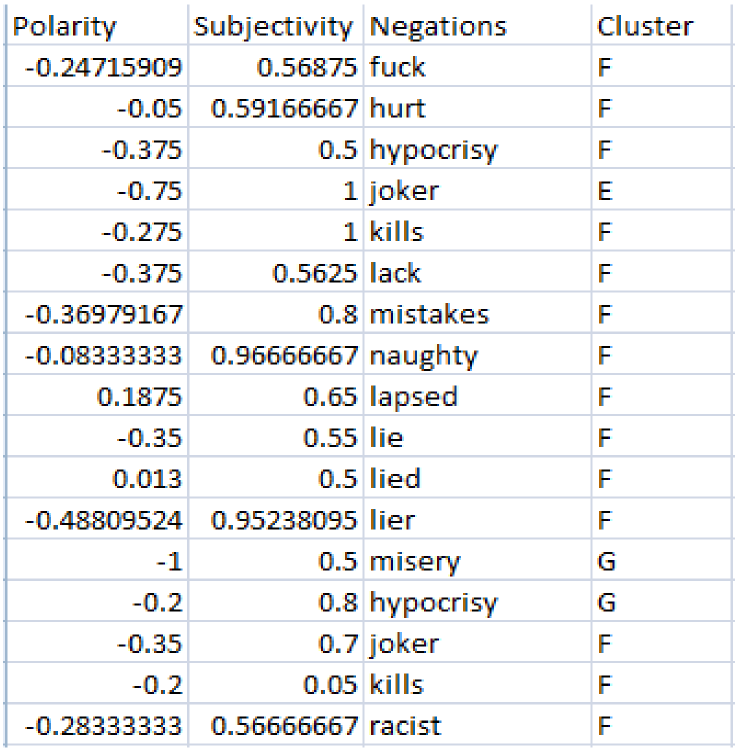
STEP2: Build Naïve Bayes classifier model on WEKA

1. I have converted the CSV files into an attribute-relation file format (ARFF) to make the dataset compatible for validation

2. Create a model of each analyzer by providing the training set file.

STEP 3: Execution of model on the test set

1. Load the test set file.
2. Execute the model on the test set.
3. Save results in the output file.



Polarity	Subjectivity	Negations	Cluster
-0.24715909	0.56875	fuck	F
-0.05	0.59166667	hurt	F
-0.375	0.5	hypocrisy	F
-0.75	1	joker	E
-0.275	1	kills	F
-0.375	0.5625	lack	F
-0.36979167	0.8	mistakes	F
-0.08333333	0.96666667	naughty	F
0.1875	0.65	lapsed	F
-0.35	0.55	lie	F
0.013	0.5	lied	F
-0.48809524	0.95238095	lier	F
-1	0.5	misery	G
-0.2	0.8	hypocrisy	G
-0.35	0.7	joker	F
-0.2	0.05	kills	F
-0.28333333	0.56666667	racist	F

Figure 10: Screen-shot of CSV file

For performance evaluation, I have to compute confusion matrix, True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN), Accuracy percentage, Root Mean Square Error (RMSE) and Mean Absolute Value.

Confusion Matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model(or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

Total Instances	Predicate class= “YES”	Predicate class= “NO”
Actual class= “YES”	True Positive	False Negative
Actual class= “No”	False Positive	True Negative

Figure 11: Confusion Matrix

In Figure 11, each row represents the instances in an actual class while each column represents the instances in a predicted class, and it can be also presented swapping rows and columns (column for the actual class, row for predicted class). To determine a confusion matrix, we have to deal with True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN).

True Positive (TP):

If the outcome of a prediction is p and the actual value is also positive, thenit is called a true positive (TP).

False Negative (FN):

If the outcome of a prediction is negative and the actual value is also positive, thenit is called a false negative (FN).

False Positive (FP):

If the outcome of a prediction is positive but the actual value is negative, then it is said to be a false positive (FP).

True Negative (TN):

If the outcome of a prediction is negative and the actual value is also negative, thenit is called a true negative (TN).

So the ACCURACY (ACC) or Classification Rate, I mentioned above, can be expressed as below:

ACC=$\frac{TruePositive+TrueNegative}{Positive+Negative} = \frac{TruePositive+TrueNegative}{TruePositive+FalseNegative+TrueNegative+FalsePositive}$

Recall:

Recall in this context is also referred to as the true positive rate. The recall is the fraction of relevant instances that have been retrieved from the total amount of relevant instances. The recall is the number of True Positives divided by the number of True Positives and the number of False Negatives. Put another way it is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate.

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

Precision:

Precision is also referred to as positive predictive value (PPV).precision is the fraction of relevant instances among the retrieved instances.Precision is the number of True Positives divided by the number of True Positives and False Positives. Put another way, it is the number of positive predictions divided by the total number of positive class values predicted. It is also called the Positive Predictive Value (PPV).

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

F1 Score:

The F1 score is a measure of a test's accuracy. The F1 Score can be written as:

2*((precision*recall) / (precision + recall)). It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall.

$$F1Score = \frac{2 * Precision * recall}{Precision + recall}$$

Accuracy Percentage:

A quick way to evaluate a set of prediction on a classification problem is by using accuracy. Classification accuracy is a ratio of the number of the correct prediction out of all predictions that were made. It is often presented as a percentage of 0% for the worst possible accuracy and 100% for the best possible accuracy.

$$Accuracy = \frac{CorrectProdictions}{TotalPredections} * 100$$

Mean Absolute Error:

Mean Absolute Error is a measure of the difference between two continuous variables. It is the difference between the measured values (Prediction) and “true” value. It gives an idea of how wrong the predictions were. The formula is given as:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

Root Mean Square Error:

The Mean Squared Error or (MSE) is much like the mean absolute error in that it provides a gross idea of the magnitude of the error.Root Mean Squared Error (RMSE) is a quadratic scoring rule that also measures the average magnitude of the error. It’s the square root of the average of squared differences between prediction and actual observation. Taking the square root of the mean squared error converts the units back to the original units of the output variable and can be meaningful for description.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

4.2.5.2.3 Result and Analysis:

I have done the Naive Bayes classification with the following features:

- Test mode: 10-fold cross-validation
- Use supervised discretization
- Numeric is taken up to 4 decimal

Summary of Naïve Bayes Classification:

Correctly Classified Instances	889
Incorrectly Classified Instances	207
Kappa statistic	0.7137
Mean absolute error	0.1979
Root mean squared error	0.3226
Relative absolute error	45.1536 %
Total Number of Instances	1096

Detailed Accuracy By Class:

TP rate	FP rate	Recall	Precision	F-measure	Accuracy
0.794	0.091	0.794	0.751	0.772	81.11%

Confusion Matrix:

a	b	c	<-- classified as
348	51	27	|a=E
24	318	47	|b=F
18	40	223	|c=G

Figure 11: Confusion Matrix Created by Naive-Bayes.

Visualization of Instances:

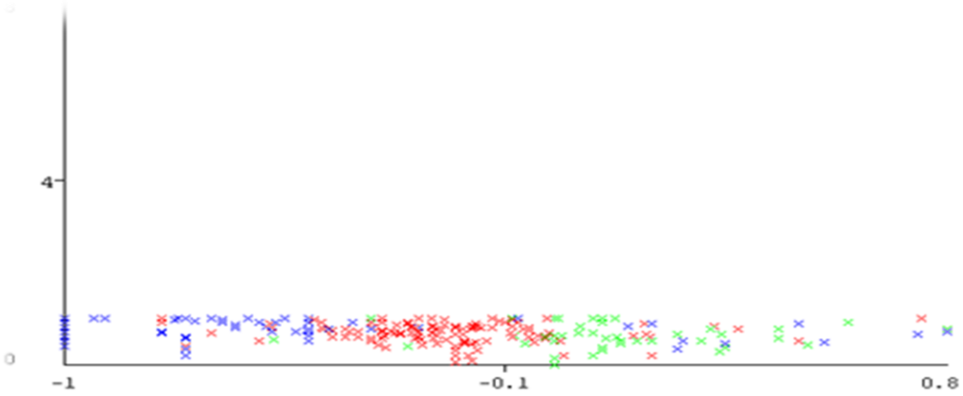


Figure 12: Visualization of instances

The x-axis denotes Polarity of Instances and Y-axis denotes Subjectivity of Instances.

Class color: EFG

4.2.6 Conclusion

Naive –Bayes classifier gives us the accuracy of 81.11%. The result shows that it is a motivating technique.Naive Bayes classifier has been used to predict the class of individual negation and the predicted class obtained from Naive Bayes is compared with the predicted class by Lexicon Based Approach. Our performance on the classification of negation may be diverted at some point due to the presence of spelling mistakes, some words used as both negative and positive sense, use of acronym etc. The results of this work serve as a partial view of the phenomenon. More research needs to be done in order to validate or invalidate these findings, using larger samples.

Chapter Five

Conclusion and Future Work

5.1 Conclusion

Twitter is a demandable microblogging service which has been built to discover what is happening at any moment in time and anywhere in the world. In the survey, we found that social media-related features can be used to predict negation of sentiment in Twitter.

I have used three models namely API based, Dictionary based model and Naive-Bayes Classification by using WEKA. This proposed system concludes the sentiments of tweets which are extracted from twitter and classify the negation during this analysis. I have also given a statistical analysis on a number of emoticons, negations, positive words, negative words in Twitter data. In this work, my results showed the accuracy, precision, recall and F1 score of the Naïve Bayes classifier. Here we got an accuracy of 81.11% by using this method. The difficulty increases with the nuance and complexity of opinions expressed in English languages. And sometimes I have found spelling mistakes, high use of acronym words, it hurt my performances. The different methods can be implemented in this project like SVM or maximum entropy model on the classification of negation over a large data set.

5.2 Future Work

❑ The sentiment analysis problem can be solved to a satisfactory label by manual training. But a fully automated system for sentiment analysis which needs to manual intervention has not been introduced yet. This is one of the main challenges in this field.

❑ In this work, I have obtained the result when analyzing tweets with TextBlob and sentiment API's, as is clearly shown in the results. In order to take the initiative to next level, I will find the patterns of emotional hypsographic parties based on Twitter. Further analysis can be done to negation in images and all types of multimedia files based on index support.

❑ Right now I have worked with only the very simplest Machine Learning Model and Lexicon Based Model; I can improve those models by adding extra information like the closeness of the word with a negation word. I could specify a window prior to the word (a window could, for example, be of 2 or 3 words) under consideration and the effect of negation may be incorporated into the model if it lies within that window. The closer the negation word is to the dictionary word whose prior polarity is to be calculated, the more it should affect the polarity. For example, if the negation is right next to the word, it may simply reverse the polarity of that word [17].

❑ Apart from this, I am currently only focusing on outer meaning of a sentence.Right now I am exploring Parts of Speech separate from the machine learning models, I can also try to incorporate POS information within our working Machine Learning models in future. So say instead of calculating a single probability for each word and emoticons. I could instead have multiple probabilities for each according to the Part of Speech the word belongs to. For example I may have P (word | object, verb), P (word | object, noun) and P (word | object, adjective). Appending POS information may change in performance significantly as they denote as a single word in the dictionary, not accordant to their meaning [13]. However, these results are for classification of reviews and may be verified for sentiment analysis on micro-blogging websites like Twitter.

❑ One more feature that is worth exploring is whether the information about relative position of the word in a tweet has any effect on the performance of the classifier. As sometimes the position of the emoticons and negation words can change the full meaning and polarity of sentences. Besides being the three classes (positive, negative and neutral) are not equal. So it affects the overall polarity [33].

❑ Last but not the least, I can attempt to model few more co-efficient and parameter and degrees in our system, which gives better accuracy. The Machine Learning classifiers fail to predict positive review due to the presence of dual-negative words.The better accuracy can be obtained with Deep learning, an emerging and growing field of research in Intelligent Systems.

References

[1] Tottie, Gunnel. 1991. Negation in English Speech and Writing: A Study in Variation.Academic Press, New York.

[2] “On the Natural History of Negative Polarity Items” Jack Hoeksema University of Groningen, http://www.let.rug.nl/hoeksema/gram-npi.pdf.

[3] Talmy Givón. 1993. English grammar: A function-based introduction. Amsterdam, Netherlands.

[4] Vapnik V. “Support Vector networks”, presented at machine learning; 1995.

[5] Payne, Thomas E. 1997. Describing Morphosyntax. Cambridge University Press, Cambridge, UK.

[6] Van derWouden, Ton. 1997. Negative Contexts: Collocation, Polarity, and Multiple Negation. Routledge, London.

[7] WendyW. Chapman, WillBridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 34(5):301–310, October.

[8] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[9] B. Pang, L. Lee and S. Vaithyanathan, “Thumbs Up? Sentiment Classification Using Machine Learning Techniques,” Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02, vol. 10, pp. 79-86, 2002.

[10] Fuchun Peng. 2003, Augmenting Naive Bayes Classifiers with Statistical Language Models.

[11] Bing Liu, Minqing Hu, and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

[12] S. Das and M. Chen, “Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web”, Management Science, vol. 53 (9), 2004.

[13]Apoorv Agarwal, BoyiXie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, “Sentiment Analysis of Twitter Data” Department of Computer Science, Columbia University, New York, 2009.

[14] R. Parikh and M. Movassate, “Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009.

[15] L. Jia, C. Yu,and W. Meng, “The effect of negation on sentiment analysis and retrieval Effectiveness,” Proceeding of the 18th ACM conference, no. c, pp. 1827-1830, 2009.

[16] Pak and P. Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010.

[17] Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[18] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volume pages 241{249, Beijing, August 2010.

[19] A. Hogenboom, P. Van Iterson, B. Heerschop, F. Frasincar and U. Kaymak, “Determining negation scope and strength in sentiment analysis,” Conference Proceedings - IEEE InternationalConference on Systems, Man, and Cybernetics, pp. 2589-2594, 2011.

[20] M. Dadvar, C. Hauff and F. D. Jong, “Scope of Negation Detection in Sentiment Analysis,” Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2011), pp. 16-20, 2011.

[21] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[22] Balakrishnan Gokulakrishnan, “Opinion Mining and Sentiment Analysis on a Twitter Data Stream”, The International Conference on Advances in ICT for Emerging Regions – ICT, 2012: 182-188.

[23] Kafuman JM. JMaxAlign : A Maximum Entropy Parallel Sentence Alignment Tool In Proceedings of COLING'12: Mumbai, 2012, p-277-88.

[24] C.C. Aggarwal and T. Abdelzaher, ”Social sensing,” in Managing and mining sensor data, Springer US, 2013, pp. 237-297.

[25] V. Singh and S. K. Dubey, “Opinion mining and analysis: A literature review”, in 5th Int. Conf. on Confluence the Next Generation Information Technology Summit (Confluence), 2014, pp. 232-239.

[26] Calvin and Johan Setiawan, “Using Text Mining to Analyze Mobile Phone Provider Service Quality (Case Study: Social Media Twitter)” International Journal of Machine Learning and Computing, Vol. 4, No. 1, February 2014.

[27] Aizhan Bizhanova, Osamu Uchida, “Product Reputation Trend Extraction from Twitter” Social Networking, 2014, 3, 196-202.

[28] Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social Media Data", IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5.

[29] Pablo Gamallo, Marcos Garcia, “Citrus: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Aug 23-24 2014, pp 171-175.

[30] Franky, O. Bojar and K. Veselovská, “Resources for Indonesian Sentiment Analysis,” The Prague Bulletin of Mathematical Linguistics, vol. 103, no. 1, pp. 21-41, 2015.

[31] B. Supriyono, “web data mining for customer's sentiment classification for Telkom speedy using twitter in Indonesian,” no. August 2015.

[32] Hao Wang, Jorge A. Castanon, Silicon Valley Laboratory, IBM, San Jose, USA, “Sentiment analysis via emoticons using twitter data” 2015 IEEE International Conference on Big Data (Big Data).

[33] Sanjana Woonna and Priyanka Giri, “Sentiment analysis of Twitter data” International Journal of Innovation and Technology, 2016.

[34] Sandip D Mali, Dr. Sachin N Deshmukh, Ashish A Bhalerao, “SentiView: A Lexicon Based Approach for Twitter Sentiment Analysis” Vol. 4, Issue 11, November 2016.

[35] Mitali Desai, Mayuri A. Mehta, Computer Engineering Department, Sarvajanik College of Engineering and Technology," Techniques for Sentiment Analysis of Twitter Data: A Comprehensive Survey", in International Conference on Computing, Communication and Automation (ICCCA 2016).

[36] Kai Yang "An Effective Hybrid Model for Opinion Mining and Sentiment Analysis" 465978-15090-3015-6/17, 2017 IEEE.

[37https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

[38]https://blog.bitext.com/polarity-topic-sentiment-analysis

[39]https://www.mathworks.com/discovery/unsupervised-learning.html

[40]https://en.wikipedia.org/wiki/List_of_emoticons

[41]https://aylien.com/text-api/