Analyzing Data of Social Media to Evaluate Customer Behavior of Companies Using Sentiment Analysis

SUBMISSION DATE: 25th March, 2018

Bushra Shourin – 14101104

Kazi Sayef Shawgat – 13201046

Serajur Reza Chowdhury – 18141011

Department of Computer Science and Engineering

Supervisor:

Hossain Arif

Assistant Professor

Department of Computer Science and Engineering

Declaration

We, hereby declare that, we conducted the thesis based on our research and results we have found. Sources of others' research we used in our thesis are mentioned in the references.

Signature of supervisor

Signature of Author

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Bushra Shourin-14101104
Kazi Sayef Shawgat-13201046
Serajur Reza Chowdhury-18141011

Abstract

In this modern age where providing and consuming services through online exchange has become a daily chore, everyone loves to participate and give opinions about the services he or she consumes. Nowadays this participation is taking place much more on social sites rather than in a complain box. Facebook is one of the most commonly used social media sites where people voice their opinions on just about everything. So many service providers take the platform to promote their services among the customers. Tele-communication sector is one of them. It is quite apparent that many Tele-communication company maintains a Facebook page or group to promote their services to the customers and get feedback from them. A large number of data is being produced in this way daily.

Telephone companies collect data of customers and often provide special offers to all or particular customers. Customers give their view about those offers on social networking sites. With the help of their opinions the offers become more realistic, attractive and in the meantime profitable. But the amount of data being produced daily is massive and growing. So it is fairly hard or improbable to go through all the data and come to a decision. It needs a special kind of procedure which has to be dynamic and efficient based on time and expense.

In this thesis, We will extract all the opinions (comments as text data) from the respective Facebook pages using the Facebook graph API provided by Facebook application and go through noise cleaning, applying algorithm and classifier to calculate the sentiment polarity to come to a decision whether the offers provided by the company are getting good feedback or being criticized. We will use Naïve Bayes classifier for our sentiment analysis.

Keywords: Opinion, Sentiment Analysis, Facebook, Facebook Graph API, Naïve Bayes classifier.

Acknowledgement

We would like to express our sincere gratitude to our honorable supervisor **Mr. Hossain Arif** for all the help and time he provided. We would also like to thank him for choosing this topic and assisting us throughout the process.

We would also like to thank **Najeefa Nikhat Chowdhury** for her guidance about collecting data from Facebook.

Table of Contents

FIGURES

Chapter 1

Introduction

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics and biometrics to systematically identify, extract, quantify and study affective states and subjective information. [1] It is vastly used in user reviews, responses in online and social media sites.

Sentiment analysis is becoming a popular study these days, mainly because of the fact that social networking sites include online users who are free to express their thoughts, feelings and impressions concerning a specific topic. Apart from their written surveys, the companies also extend their customer satisfaction analysis through the web, in order to gather a large amount of data. Sentiment analysis has also made it possible to analyze the mood of people. It can help us to decide the positive, negative or neutral views of a person based on his attitude on a given topic. Previously, it was used for lexical or syntax feature extraction, assigning a polarity label to each document or text unit.

The Sentiment Analysis is based not only on the negative or positive polarity of words and concepts, but also on the syntactical tree of the sentence being analyzed. The system tries to read between lines, identifying idiomatic or colloquial expressions, giving interpretation to negations, modifying polarity of words basing on the related adverbs, adjectives, conjunctions or verbs, taking in account specific functional-logic complements [3].

Chapter 2

Literature Review

2.1 Sentiment Analysis

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics and biometrics to systematically identify, extract, quantify and study affective states and subjective information. It is vastly used in user reviews, responses in online and social media sites.

Sentiment analysis is becoming a popular study these days, mainly because of the fact that social networking sites include online users who are free to express their thoughts, feelings and impressions concerning a specific topic. Online reviews given by customers in the social networking sites are very important to both customers and service providers. With rapid internet people typically search for things in the internet. In fact, nowadays, any kind of marketing business is currently immersing to the new trends of businesses. Apart from their written surveys, the companies also extend their customer satisfaction analysis through the web, in order to gather a large amount of data.

Sentiment analysis has also made it possible to analyze the moods of a person. It can help us to decide the positive, negative or neutral views of a person based on his attitude on a given topic. Previously, it was used for lexical or syntax feature extraction, assigning a polarity label to each document or text unit. These days, social networking sites like Twitter show the influence that surroundings have on online users [1].The Sentiment Analysis is based not only on the negative or positive polarity of words and concepts, but also on the syntactical tree of the sentence being analyzed. The system tries to read between lines, identifying idiomatic or colloquial expressions, giving interpretation to negations, modifying polarity of words basing on the related adverbs, adjectives, conjunctions or verbs, taking in account specific functional-logic complements [3].

A features dependent technique for opinion mining as well as classification was suggested by Balahur and Montoyo, which proposed a features driven opinion summarization technique wherein

the term Idquo driven rdquo explained the notion for detailing the look. The suggested technique enhanced baseline and a discussion on the technique's strong as well as weak points was suggested.

Architecture and the main components of Olympus Matini (OM) system were described by Jin et al. The new method's evaluation is on the basis of processing online product reviews from Amazon as well as other datasets. Merchants selling items on the Internet request consumers to express their opinions as well as hands-on experience regarding the particular product bought making it difficult for consumers to read and make informed decisions. The Opinion Miner system in this work mines a product's customer reviews and extracts detailed product entities where reviewers convey opinions. Opinion expressions are detected and orientations for every recognized product entity are classified as positive or negative. Differing from earlier methods that used rule-based or statistical methods, a new machine learning method built under the model of lexicalized Hidden Markov Models was proposed.

The rapid growth of computer based high-throughput method has offered unparalleled opportunities for humans to extend capacities in production, services, communications as well as research. In the meantime, huge amounts of high dimensional data are gathered to challenge amazing data mining methods. Features selection is an important stage in data mining applications that can efficiently decrease data dimensionality through removal of non-relevant attributes. In the previous few decades, researches have formulated huge quantities of features selection protocols. The protocols are formulated for serving various purposes, of various models and have their own benefits as well as shortcomings. Though there have been exhaustive efforts in reviewing already present features selection protocols as far as is known, there is no dedicated archive which gathers representative features selection for facilitating the comparison as well as joint study. For filling this gap, Zhao et al. [4] presented a features selection archive that was formulated for collecting the most famous protocols which have been formulated in the features selection research for serving as a platform to facilitate their application, comparison as well as joint study. The archive also efficiently helps research scholars in achieving more dependable evaluations in the procedure of formulating novel features selection protocols.

Systemic analyses frameworks utilizing Korean Twitter data for mining temporal as well as spatial trends of brand images was proven by Cho et al. [4] Publicly available Korean morpheme analyzer

analyzed Korean tweets grammatically, and built Korean polarity dictionaries possessing a noun, adjective, verb, and/or root for analyzing every tweet's sentiment. Sentiment classification is carried out by a SVM as well as multi-nominal Naive Bayes classifier.

2.2 Big Data

Big data is extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. The term "Big Data" used in predictive analysis, user behavior analysis or certain advanced data analytic method that used to extract value from data. Data set grows rapidly because information-sensing internet of things devices such as mobile devices, aerial, software logs, cameras, microphones etc. It is very important is machine learning, digital footprint and business intelligence. Sunil B. Mane [27] proposed that Twitter contains huge amount of useful data in form of tweets, which can be used for commercial benefits. Using sentimental analysis over Hadoop cluster as data processing agent in real time is discussed. The paper concludes with the fact that emoticons and hashtags used in the tweets can be very useful asset in sentimental analysis. However, in this, sentimental analysis with emoticons limits accuracy of analysis to only *3/4th* of a percentage. Manisha Sahane [28] focused the research towards use of modern data storage and processing technique x2013; Hadoop with prime focus on Map Reduce technique. She proposed that present data generation sources have increased drastically and thus there is a shear need of modern and more efficient data processing tools. Conclusion are made that Hadoop having HDFS as voluminous data storage cell and Map Reduce as a factor which process this huge data efficiently. Lokmanyathilak [29] made research on the computational infra of quick opinion mining. He targeted the customers switching from one free-software to another due to bugs or failure. The proposed system used twitter as source for feedback mining and applying fast-feedback opinion mining to cater customer's opinion on the software. Conclusion says that instead of using default failure reporting system, the proposed system gave about 80% accurate results.

Big data, sentiment analysis, Naïve Bayes' classifier are playing a very important part in today's research field. Because of huge amount of data produced in everyday, big data is becoming essential for detecting customer behavior. It is also important for brands to get the customer

information. They can store information so that they can detect customer views about their services. Sentiment analysis used to detect customer behavior. Naïve Bayes' classifier is used to get the polarity of views. It divides every data into three parts positive, negative and neutral. Using these three things we can detect what is true or false, what is fake or real, what is positive or negative. Within a lot of data, we can easily distinguish between positive and negative.

2.3 Applications

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level. There are many established methods for sentiment analysis at the sentence and paragraph level. In [10], the authors discussed the application of support vector machines in sentiment analysis with diverse information source. In [11], the authors applied minimum cuts in graphs to extract the subjective portion of texts they were studying and used machine learning methods to perform sentiment analysis on those snippets of texts only. In [12], the authors discussed categorizing texts into polar and neutral first before determining whether a positive or negative sentiment is expressed through the text. However, in [7], the authors operate on the premise that little neutrality exists in online texts. In [13], the authors developed techniques that algorithmically identify large number (hundreds) of adjectives, each with an assigned score of polarity, from around a dozens of seed adjectives. Their methods expand two clusters of adjectives (positive and negative word groups) by recursively querying the synonyms and antonyms from WordNet. Since recursive search quickly connects words from the two clusters, they implemented several precaution measures such as assigning weights which decrease exponentially as the number of hops increases. This confirms that the algorithm-generated adjectives are highly accurate by comparing them to the results of manually picked word lists It is worth pointing out that this work uses Lydia as the backbone to process large amount of news and blogs. In [14], the authors provided a good survey of various techniques developed in online sentiment analysis. It covers concept of emotion in written text (appraisal theory), various methodologies which can be broadly divided into two groups: (i) symbolic techniques that focuses on the force and direction of individual words (the so-called "bagof words" approach), and (ii) machine learning techniques that characterizes vocabularies in context. Based on the survey, the authors found that symbolic techniques achieve accuracy lower than 80% and are generally poorer

than machine learning methods on movie review sentiment analysis. Among the machine learning methods, they considered three supervised approaches: Support Vector Machine (SVM), Naive Bayes Multinomial (NBM), and maximum Entropy (Maxent). They found that all of them deliver comparable results on various feature extraction (unigrams, bigrams, etc) with high accuracy at 80%~87%. Another significant effort for sentiment classification on Twitter data is conducted by [2]. The authors use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like retweet, hash tags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. In [15], the authors perform sentiment analysis on feedback data from Global Support Services survey. One aim of their study is to analyze the role of linguistic features like POS tags. They perform extensive feature analysis and feature selection and demonstrate that abstract linguistic analysis features contributes to the classifier accuracy.In [16], the authors use distant learning to acquire sentiment data. They use tweets ending in positive emoticons like ":)" ":-)" as positive and negative emoticons like ":(" ":-(" as negative. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM), and they report SVM outperforms other classifiers. In terms of feature space, they try a Unigram, Bigram model in conjunction with parts-of-speech (POS) features. They note that the unigram model outperforms all other models. Specifically, bigrams and POS features do not help. In [17], the authors take a naive approach to collect and classify 300000 tweets into three categories: (i) tweets queried with emoticon queries such as ":-)", ":)", "=)" indicate happiness and positive emotion (ii) tweets with ":(", ":(", "=(", ";(" implies dislike or negative opinions, and (iii) tweets posted by newspaper accounts such as "New York Times" are considered objective or neutral. This serves as the training set for Naive Bayes Multinomial (NBM), which they found to be superior to Support Vector Machine (SVM) and Conditional Random Field (CRF) as the classifier to unigrams, bigrams, and trigrams. The result indicates that bigrams provides the best accuracy.

However, after a thorough investigation in the related scientific literature, we came up with the result that there is not any sentiment analysis and recognition using the Naive Bayes classifier for language learning in Facebook. Most of the aforementioned approaches, however, are primarily based on ngram models. Moreover, the data they use for training and testing is collected by search

queries and is therefore biased. In contrast, we present features that achieve a significant gain over a unigram baseline. In addition, we explore a different method of data representation and report significant improvement over the unigram models. Our data are a random sample of streaming Facebook statuses unlike data collected by using specific queries. The size of our hand-labeled data allows us to perform cross validation experiments and check for the variance in performance of the classifier across folds.

Chapter 3

Methodology

Twitter and Facebook are the most used social network sites. We are targeting Facebook for a special purpose. Data extraction from twitter and Facebook is different. In Twitter system collect raw data using hashtags, like #beautiful or #Disappointed, and use the data as a corpus to be fed upon implementing the classifying method. It has one downside. Every Twitter data is restricted to 280 characters in length. For this Twitter users use so many abbreviations and fragmented expressions. But unlike Twitter, Facebook has 5000 characters for every status updates and comments. [6] For this it is very much expected and convenient to get a full organized and clear sentences. Then again the number of Facebook user is much larger than Twitter which helps to get a good number of data to create a corpus. As a result, the decision about the product or service would get more précised and interesting.

As we mentioned above that we will collect our sample data as a corpus and store them in an excel file, this process will be done using the Facebook Graph API and with the help of a python script. This script can extract all the post and comments related to the posts respectively and will create different excel sheet as output. The script will also be able to extract the number of reactions every post gets from the users which will be stored in the same excel sheet.

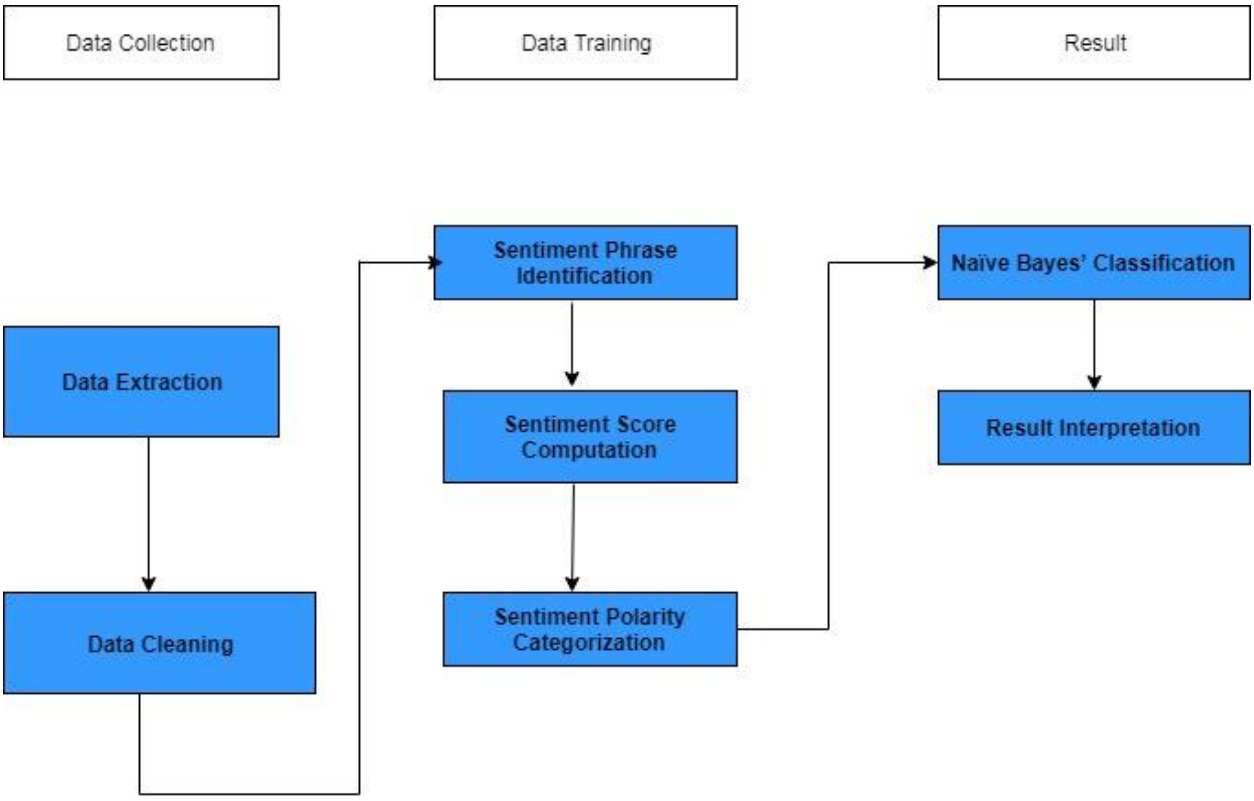


Figure 3.1: Flow Chart of Research Phases

3.1 Facebook Graph API

In order to use the Facebook Graph API effectively, we need to have an account on Facebook and create a developers account in the Facebook for developers' section. (Fig: 3.2) Then Facebook will allow you to extract all the public data from any public groups, pages or even from the public profile. In this case we need a secret application token which is given by the Facebook application. By this token Facebook get confirmation about the authenticity of the users. But this token gets expired within a very short time approximately within an hour. Then again another approach is to use the concatenation form of Facebook app ID and Facebook Secret ID to use it as the token. It never expires.

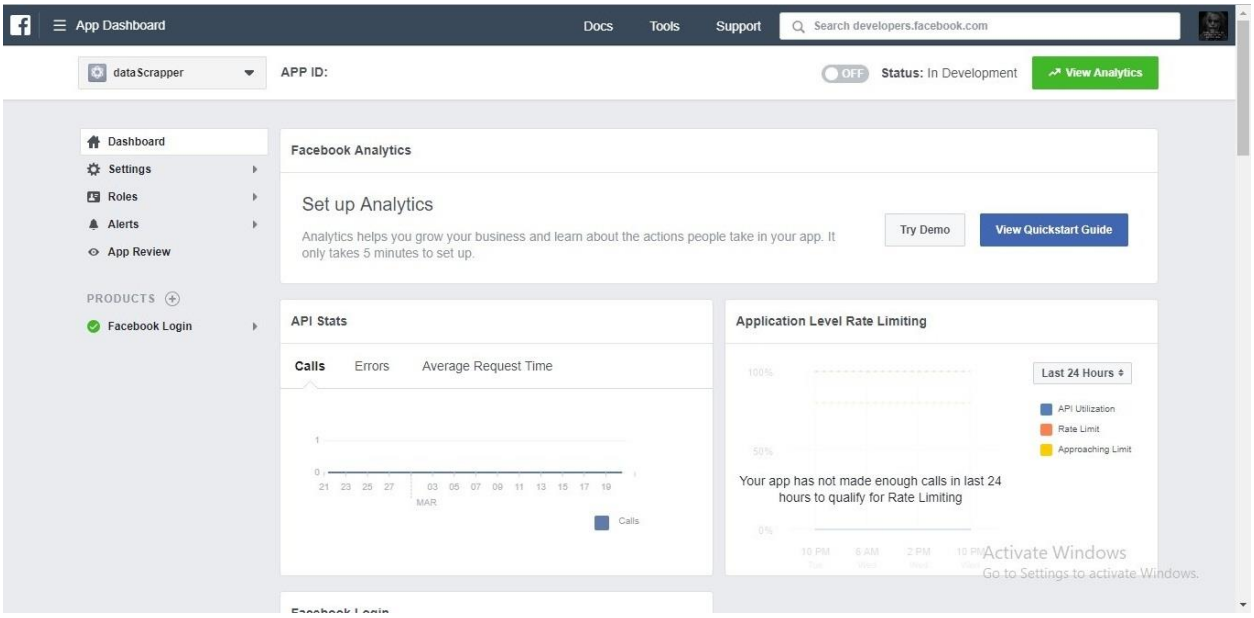


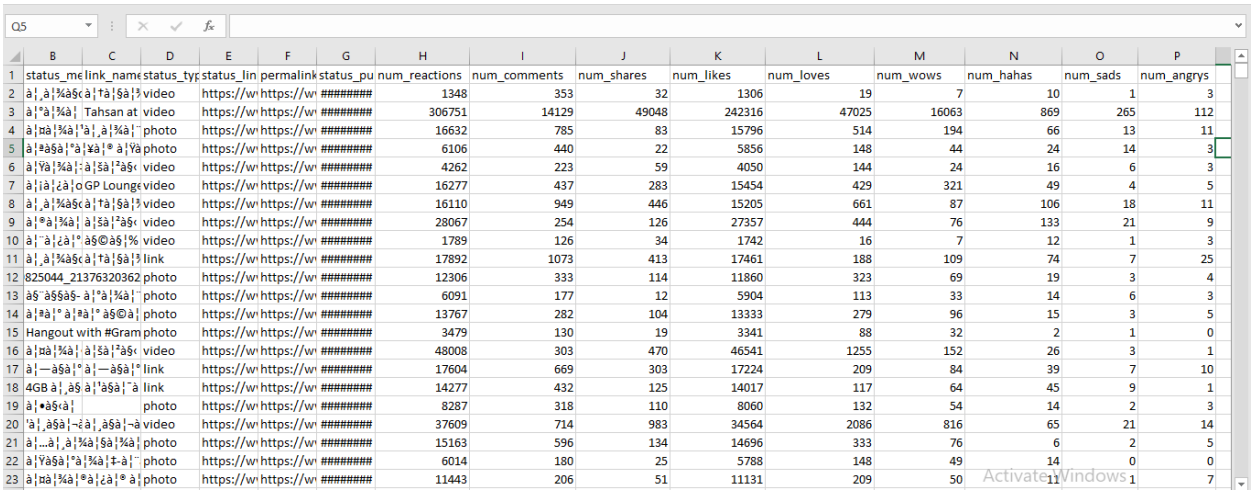
Figure 3.2: Facebook Developers Tool.

3.2 Python Script

After getting the authentication from the Facebook to use the Facebook Graph API we are ready to extract our desired data set from the Facebook. In this case we will use Python libraries and Python script to extract data. [8] Our script will work fine in Python 2.7 and above. We need some special featured libraries in our script such as Jason, Naïve Bayes Classifier, Indic, time and csv.

3.3 Dataset

After running the python Script our data will be stored in an excel file automatically. The post.py script will extract all the posts posted by the public and the comments.py will extract all the comments with respect to the posts id in an another excel file along with the counts of the reaction of that post. (Fig: 3.2)



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	status_me	link_name	status_typ	status_lin	permalink	status_pu	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
2	à¦‚à¦¾à§cà¦†à¦§		video	https://w	https://w	########	1348	353	32	1306	19	7	10	1	3
3	à¦°à¦¾ Tahsan at		video	https://w	https://w	########	306751	14129	49048	242316	47025	16063	869	265	112
4	à¦¤à¦¾à¦'à¦‚à¦¾à¦¨		photo	https://w	https://w	########	16632	785	83	15796	514	194	66	13	11
5	à¦ªà§à¦°à¦¥		photo	https://w	https://w	########	6106	440	22	5856	148	44	24	14	3
6	à¦Ÿà¦¾à¦:à¦§à¦²à§(		video	https://w	https://w	########	4262	223	59	4050	144	24	16	6	3
7	à¦ià¦¿à¦oGP Lounge		video	https://w	https://w	########	16277	437	283	15454	429	321	49	4	5
8	à¦‚à¦¾à§cà¦†à¦§		video	https://w	https://w	########	16110	949	446	15205	661	87	106	18	11
9	à¦°à¦¾à¦ à¦§à¦²à§(		video	https://w	https://w	########	28067	254	126	27357	444	76	133	21	9
10	à¦¨à¦¿à¦°à§©à§%		video	https://w	https://w	########	1789	126	34	1742	16	7	12	1	3
11	à¦‚à¦¾à§cà¦†à¦§		link	https://w	https://w	########	17892	1073	413	17461	188	109	74	7	25
12	825044_21376320362		photo	https://w	https://w	########	12306	333	114	11860	323	69	19	3	4
13	à§¨à§§à§- à¦°à¦¾à¦		photo	https://w	https://w	########	6091	177	12	5904	113	33	14	6	3
14	à¦¤à¦° à¦¤à¦° à§©à¦		photo	https://w	https://w	########	13767	282	104	13333	279	96	15	3	5
15	Hangout with #Gram		photo	https://w	https://w	########	3479	130	19	3341	88	32	2	1	0
16	à¦¤à¦¾à¦ à§à¦²à§(		video	https://w	https://w	########	48008	303	470	46541	1255	152	26	3	1
17	à¦—à§à¦°à¦—à§à¦°		link	https://w	https://w	########	17604	669	303	17224	209	84	39	7	10
18	4GB à¦‚à§à¦'à§à¦¯à		link	https://w	https://w	########	14277	432	125	14017	117	64	45	9	1
19	à¦•à§‹à¦		photo	https://w	https://w	########	8287	318	110	8060	132	54	14	2	3
20	'à¦‚à§à¦¬¿à¦‚à§à¦¬à		video	https://w	https://w	########	37609	714	983	34564	2086	816	65	21	14
21	à¦…à¦‚à¦¾à¦¾à§à¦¾à		photo	https://w	https://w	########	15163	596	134	14696	333	76	6	2	5
22	à¦Ÿà§à¦°à¦¾à¦‡à-à¦¨		photo	https://w	https://w	########	6014	180	25	5788	148	49	14	0	0
23	à¦¤à¦¾à¦°à¦¿à¦¿à¦° à¦		photo	https://w	https://w	########	11443	206	51	11131	209	50	11	1	7

Figure 3.3: Reactions each of the posts gets.

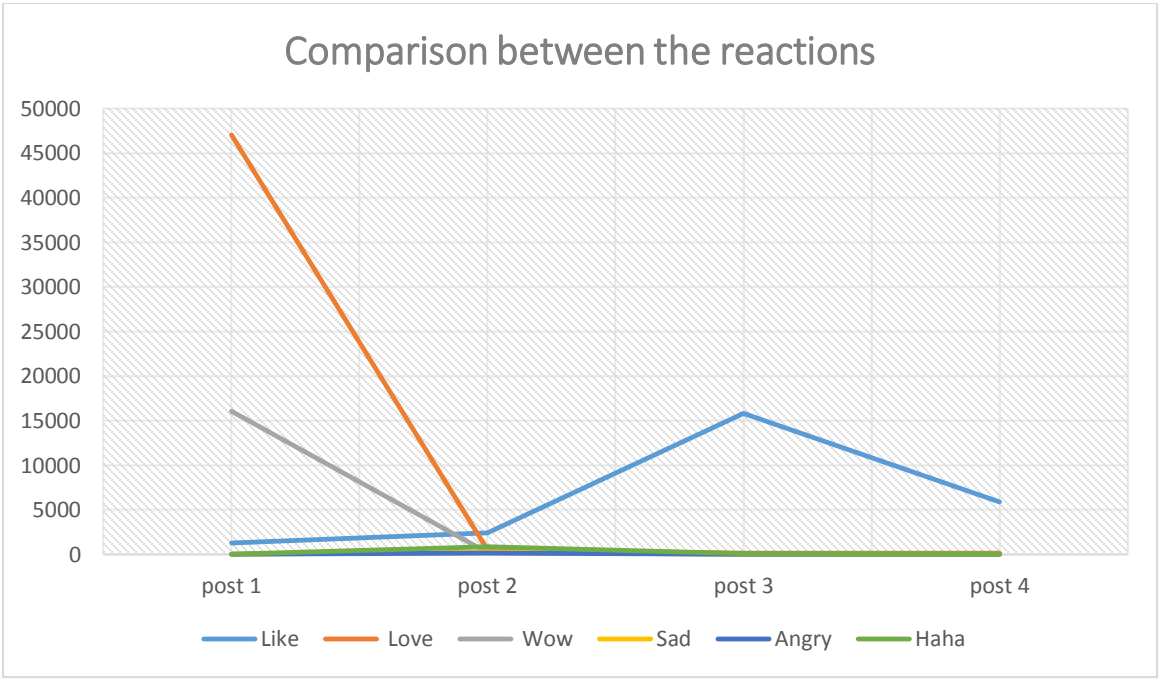


Figure 3.4: A graphical representation of the reactions of the posts.

3.4 Noise Reduction

This data set contains a lot of irrelevant words and sentences which consider as noise in big data. So we need to remove all the noises from the data. To refine our data, we use the Microsoft Excel tool and its cleaning process by calling many built in method of it. After that we convert the excel file into simple text file to split all the words into token using lexical analyzer tool.

3.5 Naïve Bayes' classifier

Naive Bayes is one of the most used techniques when it comes to text classification problems which involve high dimensional training data sets. It is based on Bayes' probability theorem. It presents less difficulty than other algorithms. It is not only fast but also very productive in terms of making predictions based on relatively small amount of currently provided data.

Even though it is based on the Bayes' theorem, it is tagged "naïve" because it makes the assumption that the occurrence of a certain feature in a class is independent of the occurrence of other features in the class regardless of any correlation of occurrences among them. For example, a fruit may be considered to be an apple if it is red, round and about 3 inches in diameter. Even if the features – color, shape and diameter depend on each other, all of these properties contribute to the decision that this fruit is an apple assuming the features to be independent of each other.

$$P(A|B) = P(B|A)\, P(A) / P(B) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Now, the conditional probability of Bayes' theorem determines the probability of an event based on the previous data of the events. In the formula mentioned above, there are two events A and B. P(A) indicates the probability of the event A and P(B) indicates the probability of the event B. So, according to the Bayes' theorem, P(A|B) or probability of A Given B is equal to P(B|A) or probability of B given A multiplied by probability of A upon probability of B. Here, A is the proposition and B is the evidence. P(B|A) indicates likelihood or how well the model predicts the data, P(A) indicates prior probability or the degree to which we believe the model accurately describes reality based data on all our prior information, P(B) indicates the Normalizing constant or the constant that makes the posterior density integrate to one and P(A|B) indicates posterior

probability or the degree to which we believe a given model accurately describes the situation given the additional available data including all our prior information. [2]

For the Naïve Bayes' classifier example, let us assume that there is a corpus of data based on comments left by Facebook users on a post. The datasets are examples of positive and negative words stored in two different classes and using the classifier algorithm we can estimate the probability of certain words' occurring or count measurement and their positive and negative sentiments. This is required as pre-classified examples to train datasets.

So, the variables in this dataset can be regarded as sentiments and the probability that a variable will occur given the evidence in the sentence can be expressed as,

$$P\ (sentiment \mid sentence) = P\ (sentence \mid sentiment)\ X\ \frac{P(sentiment)}{P(sentence)} \quad \text{.....(2)}$$

We can assume the words in a sentence as tokens and P (sentence | sentiment) as a result of P (token | sentiment) across all the words in a sentence. This will take into account the number of times a word has occurred in a sentence.

P(token|sentiment)=count(this token in class)+1/count(all tokens in class)+count (all tokens)(3)

The addition of 1 in the calculation above is called Add-One Smoothing which is used to eradicate any possibility of multiplication with zero; that means avoiding a word occurring zero times or to indicate that a word has occurred at least once more than the value presented in the training data.

Thus, the classifier firstly calculates the prior probability that is the probability of a word being positive or negative prior any trained data based on the number of positive and negative word examples. Then for each class, the tokens are multiplied with the likelihood of each word being in that class. After that the final result is measured and the highest scoring class is returned as a polarity which determines if the positivity or the negativity of the comment.

3.6 TRAINING DATASET

In this study, we emphasize on Naïve Bayes classifier to determine either a Facebook status has positive or negative feedback on the basis of its comments. First, we targeted a status and extract all the comments. Later on we prepared these comments for our training data set and test data set.

We approximately took 1000 comments from the status and trained up to 50 percent of the comments. We manually labeled the data into negative and positive polarity. We made two arrays for predictive positive words and predictive negative words. Then we count the frequency of all these positive and negative words in the sentences. According to the occurrence of a predictive positive word for the true positive sentences and in the true negative sentences we use the mean statistical formula to find a weight for each word. For this the training data set will work more efficiently. For instance, LIKE is a predictive positive word. But it may not always be in the true positive sentence. Assume that occurrence of "LIKE" in the true positive sentences are 50 times and again in the true negative sentences in 20 times. So the probability of this word for being a true positive word is 71%. So it is more likely that if LIKE exist in a sentence there is a chance that the sentence is 71 times positive within 100 times.

Sample comments of positive polarity	**Sample comments of negative polarity**
Like the offer, very effective.	This offer is not satisfactory at all.
The offer is good for the students.	The price is very high, can't afford.
Compare to others, this is a good deal.	This is so are too overrated.

Table 3.1: Sample comments of negative and positive polarity.

As we are extracting data from the Facebook we used emoticon feature to determine the polarity of sentence. There are many kinds of emoticons are used by the users. Some are used to express appreciation; some are for the anger etc. We also made two different arrays for distinguishing of its polarity and consider the above mentioned method to determine each emoticon's weight. Here are the set of emoticons we used in our analysis, [2]

Polarity	Emoticons
Positive	☺, :-p, :], ;-), :b, :-j, 8-), <3, ^_^
Negative	☹, :'(, :c, ;-(, -_-, :-@, :-/, :[

Table 3.2: Sample data for emoticons.

We trained gradually 50,100,150 and 200 datasets and test the rest amount of data respectively for finding a ratio of our result. We did because if we get a constant ratio about our positive and negative polarity we can be sure that the analysis is worked.

Chapter 4

Results

4.1 Experimental Results

We used three different methods for evaluating the performance of our sentiment analysis. Accuracy, Precision and Recall, these three methods are very effective for sentiment analysis. Accuracy will determine the overall result of our analysis but it may not always be as precise as we expect it to be. So we tried two other different methods as well. Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives. This is often at odds with recall, as an easy way to improve precision is to decrease recall. Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Improving recall can often decrease precision because it gets increasingly harder to be precise as the sample space increases. [9]

First we initialized our dataset with the positive, negative, true positive and true negative into a table to find out the accuracy, precision and recall percentage for both positive and negative set of words.

	True positive words (Heavy weight positive words)	True negative words (Heavy weight negative words)
Positive words(predicted)	a	b
Negative words(Predicted)	c	d

Table 4.1: A confusion table for finding Accuracy, Precision and Recall

Accuracy = $\frac{a+d}{a+b+c+d}$

Recall (positive) = $\frac{a}{a+c}$ Recall (negative) = $\frac{d}{b+d}$

Precision (positive) = $\frac{a}{a+b}$ Precision (negative) = $\frac{d}{c+d}$

With the help of these formula we will calculate all the percentages of the methods and we will compare the results.

The overall results of the three are indicated in the below tables with different numbers of training dataset.

Number of Trained Dataset	Accuracy (%)
50	59.6
100	66.8
150	69.2
200	72.9

Table 4.2: Accuracy result on test dataset

Number of Trained Dataset	Precision for positive corpus (%)
50	56.6
100	59.2
150	64.3
200	66.8

Table 4.3: Precision result for Positive Corpus on Test Datasets

Number of Trained Dataset	Precision for negative corpus (%)
50	55.6
100	63.2
150	66.3
200	70.8

Table 4.4: Precision result for Negative Corpus on Test Datasets

Number of Trained Dataset	Recall for positive corpus (%)
50	44.6

100	58.2
150	66.3
200	74.8

Table 4.5: Recall result for Positive Corpus on Test Datasets

Number of Trained Dataset	Recall for negative corpus (%)
50	69.1
100	66.1
150	68.6
200	73.2

Table 4.6: Recall result for Negative Corpus on Test Datasets

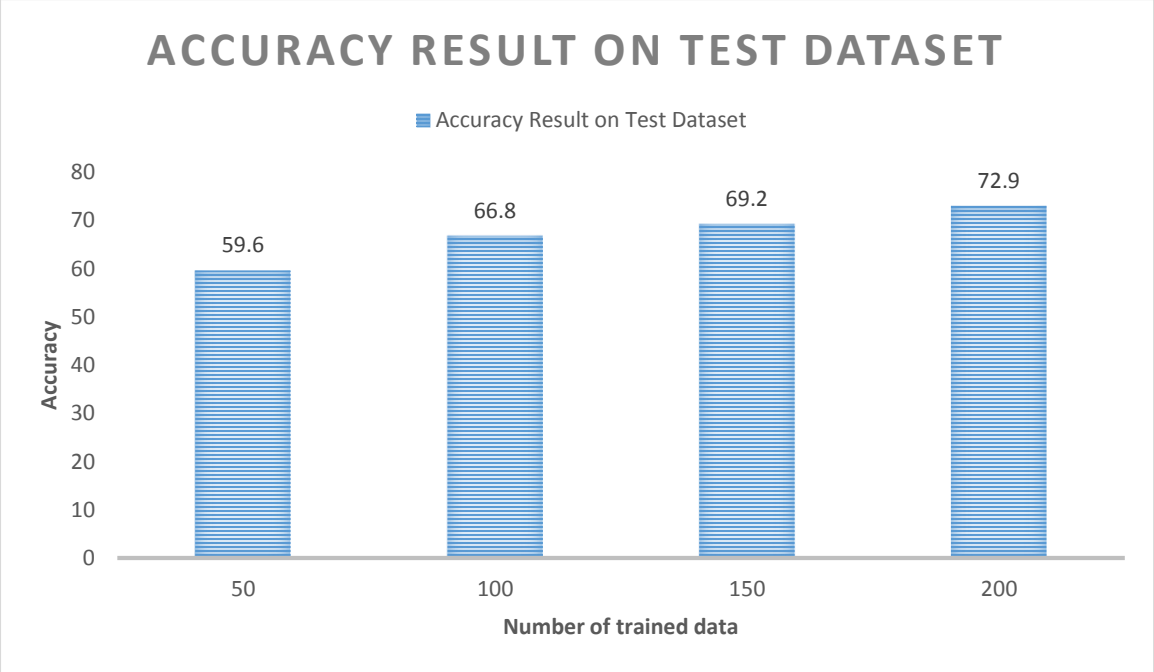


Figure 4.1: Accuracy result on test dataset

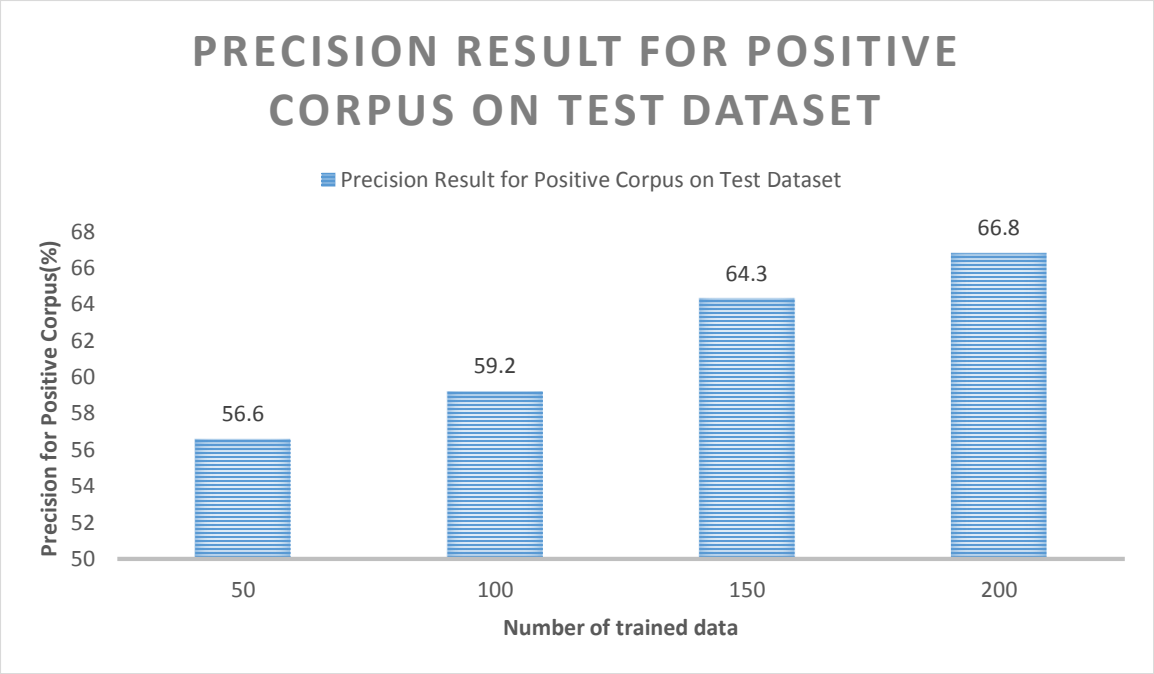


Figure 4.2: Precision result for Positive Corpus on Test Datasets

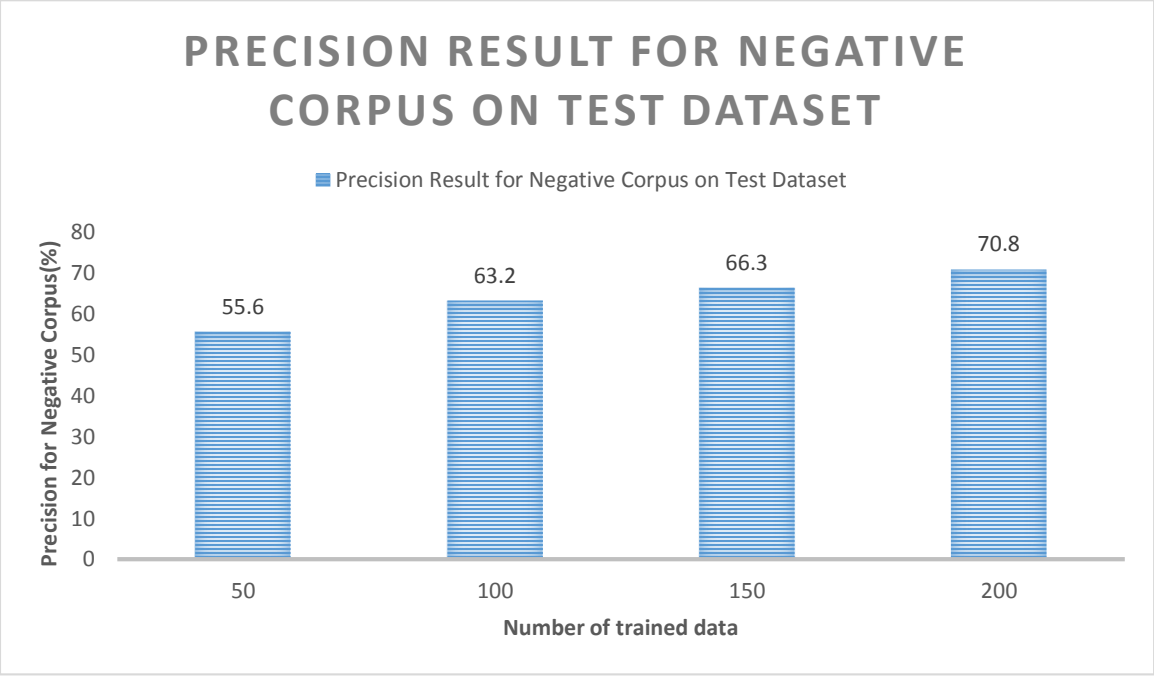


Figure 4.3: Precision result for Negative Corpus on Test Datasets

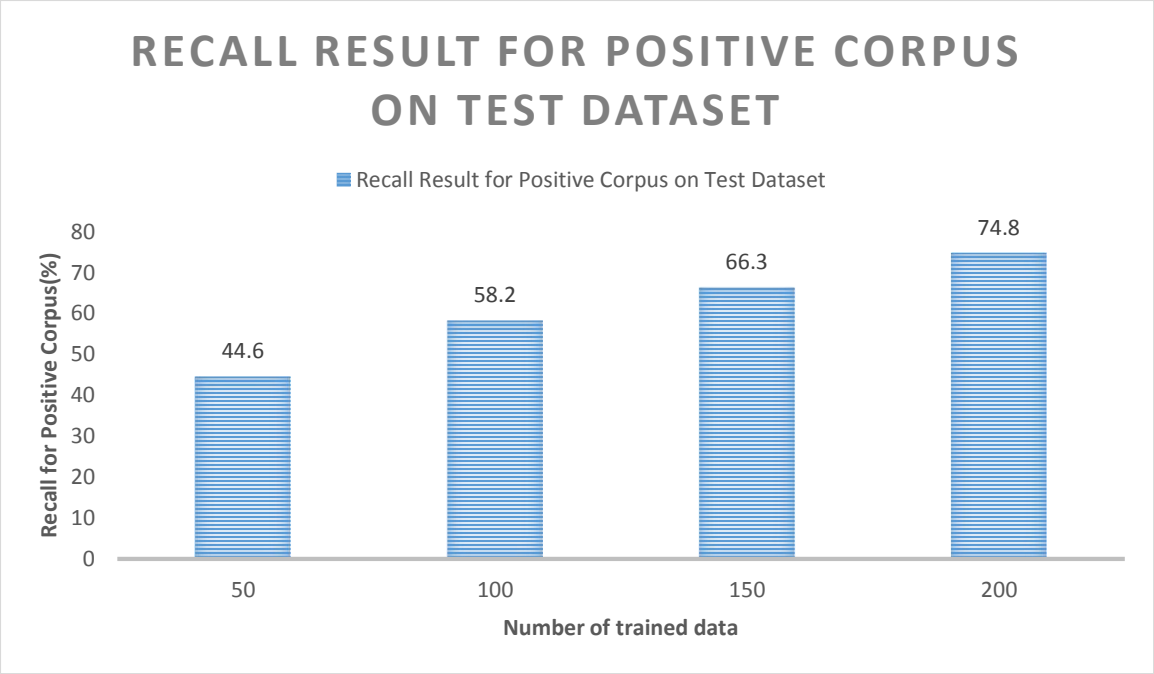


Figure 4.4: Recall result for Positive Corpus on Test Datasets

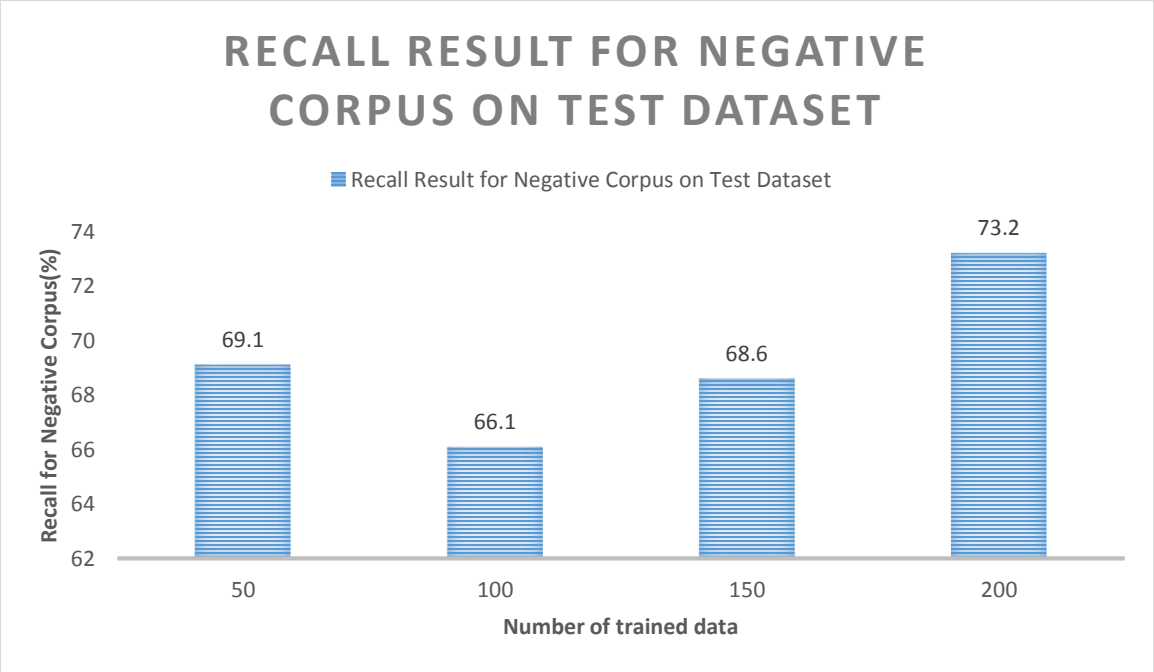


Figure 4.5: Recall result for Negative Corpus on Test Datasets

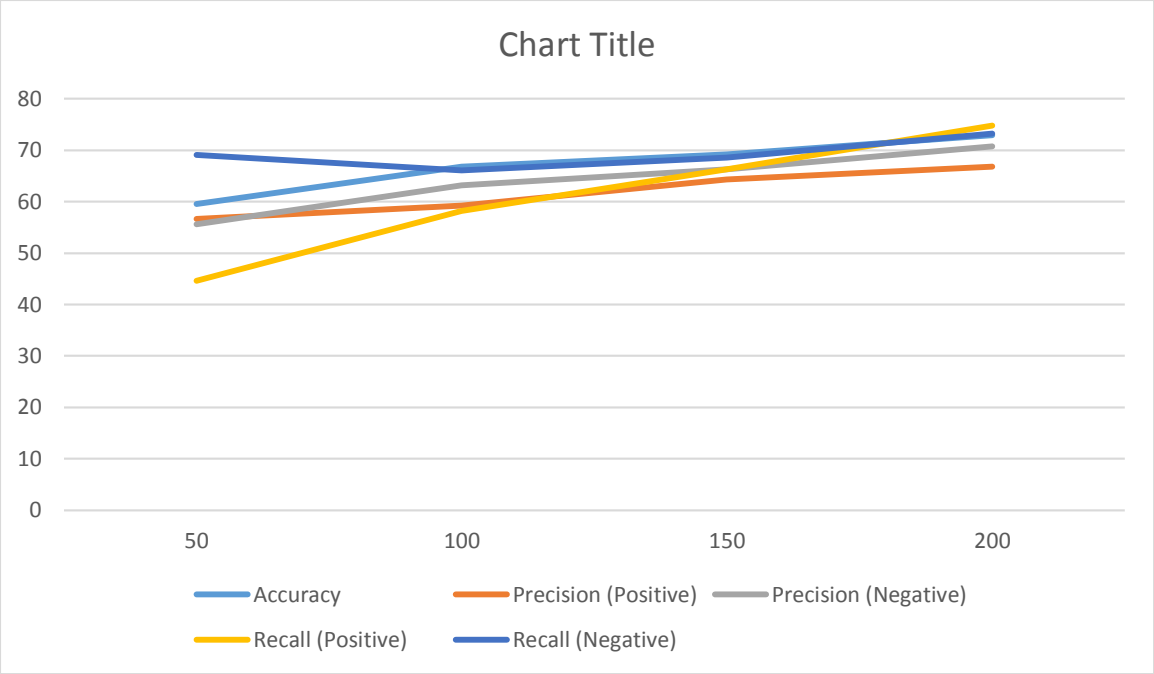


Figure 4.6: Comparison between Accuracy, Precision and Recall

4.2 Discussion

The result we got from our project is satisfactory. We expected nearly the same percentages when we cross checked manually. But we are assuming that the efficiency can be enhanced with some more algorithms and comparison. We are planning it for our future work.

In this thesis we applied an efficient and accurate but uncomplicated way to extract opinions of telecommunication users from Facebook and identify whether they accurately presented the polarity. With these results we can analyze any company's position in the market. This will be our future work.

4.3 Motivation

It's vital for business owners to pay attention to customers' feedback about their services. It's also essential for businesses determine how much of word of mouth can be counted as asset or liability to their brands' reputation. Through sentiment analysis it would be much more efficient to find out how customers feel about services, products, offers, events and even the people who are the faces of the business. Knowing what the customers are already thinking after every social media post,

business owners can form and reform their marketing strategies accordingly. It can help them obtain valuable context on how to respond to the feedback and how to approach the customers for the next service. Ways how sentiment analysis on customers' comments can uplift business situations are discussed below:

Come up with Business Solutions Beforehand:

When a news of offer or service is presented to customers on social media, business leaders can obtain comprehensive information using the positive, negative and neutral comments that will assist them to further evaluate, make reports and come up with adaptable solutions. It's very beneficial when it comes to benchmarking competitors and markets. Moreover, Sentiment analysis can also help businesses to analyze how the latest service is considered among their customers and get a general idea of which demographic segment generates the most interest for the business.

Useful for Measuring ROI of Marketing Campaign:

Simple calculation of number of likes, comments or followers do not give the real picture in terms of the success of a marketing campaign. Through sentiment analysis business owners can combine qualitative and quantitative measurements and measure the real ROI rate of the marketing campaign using the positive or negative discussions of the customers.

Boost Customer Service:

Sentiment analysis is a highly effective technique that helps companies to reach out to their customers before any negative feeling about a certain service or company's reputation spreads wide. So the companies can even turn a bad customer experience into a positive one by providing satisfying service in such cases.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The proposed system is able to collect useful information from the social networking websites and efficiently perform sentiment analysis on the data. The data we used were comments left on posts from a certain page on Facebook. Using Naïve Bayes' classifier, we were able to analyze our data with accuracy, precision and recall. Through the uncomplicated machine learning algorithm of this classifier we could classify our text data according to the sentiment polarities of the user opinions. It successfully predicted user opinions towards an offer or service of a mobile phone operator. Thus we can conclude that Naïve Bayes' classifier can be applied to effectively analyze user opinions on Facebook or any other social media sites. The operators can use this system to ensure quality services to the customers.

5.2 Future Works

Our future plan for this analysis is to study on this matter in depth and apply other sentiment analyzing algorithms such as K-nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest etc. to crosscheck and compare our results with Naïve Bayes' classifier. We would also like to come up with a new algorithm taking advantage of more than one algorithm that would further provide more accuracy for user opinions and their behavior towards certain offers or services on a post. To add more, we plan to apply a new technique on comments that include Bangla fonts used by the Bangla speaking population and incorporate them in our analysis since we are also focusing on the customer opinions of Bangladeshi mobile phone operating companies.

References

1. M. Kumar, Dr. A. Bala, "Analyzing Twitter Sentiments Through Big Data", 2016 International Conference on Computing for Sustainable Global Development (INDIACom).
2. A. Goel, J. Gautam, S. Kumar, "Real Time Sentiment Analysis of Tweets Using Naive Bayes", 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.
3. F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, T. By, "Sentiment Analysis on Social Media", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 919-926, 2012.
4. Y. Wu, F. Ren, "Learning Sentimental Influence in Twitter", International Conference on Future Computer Sciences and Application, 2011.
5. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment analysis of Twitter data", LSM '11 Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, pp. 30-38, 2011.
6. T. Mullen, N. Collier, "Sentiment analysis using support vector machines with diverse information sources", Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 412-418, 2004
7. B. Pang, L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts", ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 271-278, 2004.
8. T. Wilson, J. Wiebe, P. Hoffman, "Recognizing contextual polarity in phrase level sentiment analysis", Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347-354, 2005.
9. N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs", Proceedings of International Conference on Weblogs and Social Media, 2007.
10. E. Boiy, P. Hens, K. Deschacht, M. - F. Moens, "Automatic Sentiment Analysis in On-line Text", Proceedings of the 11th International Conference on Electronic Publishing, pp. 349-360, 2007.

11. M. Gamon, “Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis”, Proceedings of the 20th international conference on Computational Linguistics, pp. 841–847, 2004.
12. A. Go, R. Bhayani, L. Huang, “Twitter sentiment classification using distant supervision”, Technical report, Stanford, 2009.
13. A. Pak, P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010.
14. A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," *2017 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore.
15. J. Li, M. Sun, "Experimental study on sentiment Classification of Chinese Review Using Machine Learning Techniques", *Journal of computer society(IEEE)*, pp. 393-400, 2007.
16. J. Smailovic, M. Gracana, N. Lavrac, M. Znidarsic, "Stream-Based Active learning for Sentiment Analysis in the financial domain", *Journal of Information Sciences (Elsevier)*, pp. 181-203, 2014.
17. A. Balahur, M. Turchi, "Comparative experiment Using Supervised Learning and machine Translation for Multilingual Sentiment Analysis", *Computer speech and Language(Elsevier)*, pp. 56-75, 2014.
18. D. Lewis, "Naive (Bayes) at forty: The independence assumption ininformation retrieval", *Journal of Machine Learning*, pp. 4-15, 1998.
19. P. Domingos, M. Pazzani, "the optimality of the simple Bayesian classifier under zero-one loss", *Machine learningJournal of Transactions on Pattern Analysis and Machine Intelligence(IEEE)*, vol. 29, no. 23, pp. 103-130, 1997.
20. S. Liu, I. Lee, "A hybrid sentiment Analysis Framework for Large Email Data International conference on Intelligent System and Knowledge Engineering", *Conference Publishing Services*, 2015.
21. B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques", *conference on Empircal methods in natural language processing in association for computational Linguisticso1*, vol. 10, pp. 79-86, 2002.

22. B. Mutlu, M. Mutlu, K. Oztoprak and E. Dogdu, "Identifying trolls and determining terror awareness level in social networks using a scalable framework," *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016, pp. 1792-17
23. Mane, Sunil B., "Real Time Sentiment Analysis of Twitter Data Using Hadoop." IJCSIT) International Journal of Computer Science and Information Technologies 5.3 (2014): pp. 3098–3100.
24. Sahane, Manisha, Sanjay Sirsat, and Razaullah Khan. "Analysis of Research Data using MapReduce Word Count Algorithm." Internl. Journal of Advanced Research in Computer and Commn. Engg 4 (2015).
25. Selvan, Lokmanyathilak Govindan Sankar, and Teng-Sheng Moh. "A framework for fast-feedback opinion mining on Twitter data streams."Collaboration Technologies and Systems (CTS), 2015 International Conference on. IEEE, 2015.
26. K. Yadav, M. Pandey and S. S. Rautaray, "A proposed framework for feedback analysis system using big data tools," *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, 2017, pp. 542-545.