Prediction of Business Trend Using Sentiment Analysis on Social Media's Data

Md Mafidul Islam
ID: 161-15-1022

Mahfujur Rahman Mehedi
ID: 161-15-1020

This report is presented as partial fulfillment of the requirement for a bachelor's degree in Computer Science and Engineering.

Supervised By
Mr. Dewan Mamun Raza
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By
Ms. Amatul Bushra Akhi
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
10th DECEMBER 2019

DECLARATION

I hereby declare that this project has been done under the supervision of **Mr. Dewan Mamun Raza, Lecturer, Department of Computer Science and Engineering (CSE), Daffodil International University.** We also declare that neither this project nor any part of this project has been submitted elsewhere for any award of any degree or diploma.

Supervised By:

Mr. Dewan Mamun Raza
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-supervised By:

Ms. Amatul Bushra Akhi
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted By:

Md Mafidul Islam
ID: 161-15-1022
Department of CSE
Daffodil International University

Mahfujur Rahman Mehedi
ID: 161-15-1020
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we are expressing our sincere thanks and gratitude to the Almighty for His endless blessing which offers us adequate strength to finish our final year project.

We grateful to have Mr. Dewan Mamun Raza and Mrs. Ammatul Busra Akhi, Department of CSE, Daffodil International University, Dhaka in this project with us. Deep Knowledge & keen interest of our supervisor and co-supervisor in the field of “Sentiment Analysis” cheer up to continue and complete this project successfully. Their patience and energetic supervision at all stages have made it possible to complete this project.

We are very grateful to Daffodil International University, Permanent Campus, for providing a natural environment that is conducive to research and quality work.

We would like to thank my classmates at Daffodil International University, who gave us valuable suggestions during this project work.

Finally, we must accept with due respect the constant support and patience of our parents, and all the individuals who are directly or indirectly involved in the successful completion of this project work.

ABSTRACT

Social media is the largest platform to share opinion. This opinion has a great effect on business field. Because this opinion analysis will tell about one thing how much positive, negative or natural the human demand on it. Currently for Business sentiment analysis research is noticeably large, inclusive of a range of supervised studying method of classification outcome and the text characteristics depiction formula and feature choice implementation and other elements effect on the classification performance is a big problem. Sentiment analysis is research natural language processing and opinion mining. In this paper deliver an ideal decision-making process for data using predictive analysis to social media data and different types of blog data. Machine learning provides a concepted approach for developing sophisticated and automatic algorithms to analyze high-dimensional and multimodal business analytical data. This study focuses on how social media data, machine learning algorithm (Naive Bayes) and classifier find out our upcoming trend.

Keywords:

Data Mining, Sentiment Analysis, Naïve Bayes, Classifier, Business Intelligence

LIST OF CONTENTS

CONTENTS **PAGE NO**

LIST OF FIGURES

FIGURES **PAGE NO**

LIST OF TABLES

TABLES **PAGE NO**

CHAPTER 1

INTRODUCTION

1.1 Introduction

Social media has huge information and typically use for social interaction and access to news, entertainment and opinion. Collected information is an important part of our daily life which can surveillance our daily activities such as ourselves, bathing, dressing, grooming, work, homemaking, and leisure. Daily we search what is new today or scrolling website for new story. Sentiment analysis can understand what we want or what types of products we like. Then that company show us related advertisements. Sentiment analysis is a natural language process which can analyze text analysis, classification and identify our opinion. It can be possible machine learning algorithm or different types of natural languages processing classifier. Every post means a special theme and their post sentences have distinctive meaning.

Each sentence describes positive, negative or neutral which express people in the social media. We need to extract this opinion. Already many projects have been done in this related field especially business, politics, sports, publications with machine learning approaches. They have been done product review, popularity and decision making by using their post, comment, hashtag etc. from social media. So, sentiment analysis is very important for natural language processing.

Current world business pattern and trend continuously changes. Businessman should update their business pattern or grow up new business which people like most. People through opinion subconsciously new business trend or what their demand in the social media. There are more than 3 billion people use social media that means 40% of people use social media. Most of the people use Facebook, twitter to express their thoughts, opinions and blogging. It is very suitable to take data from social media and research with this data.

Although there are many projects have been done with market analysis, sports, business, product reviews, politics but they did not work with find out business trend in the world. We focus on business decision making approach and find out business trend. Many companies offer idea competition or do not know where they invest. This paper is a good solution for them. We research their post how much positive or negative or neutral and measure the frequency of words how much relation with business trend words. Which brand, product people like and why they like and their review is very important to generate new business idea. We find out this type of works word frequency and research with them.

1.2 Motivation

Business intelligence is a popular method of big companies. Their profit gets huge difference a tiny mistake or right decision. When a company bring new product and get popularity, then their related company face some problems. Customer demand on new products such as mobile company apple and Samsung use fingerprint but apple is little bit popular than Samsung. Because they bring it first. Thus, they attract people and can go to top ranking company.

Sometimes we can see in the newspaper or online portal which feature they are bringing in the market. This newspaper can assume how much sell this product watching their feature and predict their approximate profit.

Then others company take action to build up related products for do not decrease popularity. This paper based on find out upcoming product feature or idea of new product using people opinion sentiment analysis. We get data easily from twitter. We motivated business intelligence and huge social media data.

1.3 Rationale of the Study

Sentiment dataset is available in the internet. But exact business-related dataset is pretty much rare and no work has been done previous thesis projects. We create business related dataset for training purpose from internet.

Our first goal is created dataset which divided by positive or negative sentences. Positive sentence denoted by 1 and negative sentence denoted by 0. We determine positive or negative sentences in a post or paragraph. Then we find out word ranges which words are more popular. That means which words use multiple and how many times it uses. We also find out their similar words. Thus, we determine that post theme. It is slightly different from other thesis paper and unique.

1.4 Research Questions

Business intelligence is a big platform and sentiment analysis is a big platform. There have some similarities. Their larger problems solved by changing a tiny section. We need to solve their particular part and connected to each other. For example, what is the business trend? Which sector emphasize and where invest money a business man? Understand the problem every smaller part them and if we need to create complex methodology, we will create it. We will answer the following problems step by step.

1.4.1 Data collection

We need two types of data, training data and test data. We create training dataset from the internet which determine positive or negative sentences. Test data collected from twitter. It is a very challenging task for us because exact business data catching is very rare.

1.4.2 Features of data

Training data divided by two categories positive and negative. Positive and negative data indicate by 1 and 0 respectively. Test data related to business which we prepared for the test. Test dataset has three parameter id, created date, text. Test data mainly text sentences.

1.4.3 The model

There are many ways to create model for example machine learning and natural language processing classifier. Which machine learning algorithm is best and how much dataset will train? And which natural language processing classifier is suitable to get necessary output.

1.5 Expected Output

Our main target is to discover business trends from social media. We use twitter data for test purpose and create training data from online. First of all we detect positive and negative sentences. Then we will pull out the frequencies of every word, how many types that words are used in that paragraph or dataset. The dataset or paragraph will be business related. This way can understand their theme and intention.

1.6 Report Layout

This report will be written in such a way that everyone can understand what is our goal and what is our working procedure. This report follows the standard project reporting template of Daffodil International University. It has five separate parts.

Chapter 1, we already discussed in this chapter all about our research motivation, rationale of the study, research question and expected outcome.

Chapter 2, the following chapter includes the background details of sentiment analysis and business intelligence procedure and also concern about the history of Sentiment identification, problem and its challenges.

Chapter 3, we briefly discussed about our research methodology and techniques where we described data collection. Dataset creation, training, how to algorithms and classifier uses on that data set and statistical analysis about research

Chapter 4, Write with the experimental result which we get output from our algorithm and classifier.

Chapter 5, finally we talk about limitations, conclusions, future works, and a summary of the research.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

Online social media is that the vital point to several businesses as a result of the shoppers post their comments and complaints within the social media in each public and personal site. Everyday Facebook post 2.5 billion pieces of content and overall use 500+ terabytes data [20]. That is huge data like, pulling, pictures, videos etc. The content of defining an appropriate journal articles is to develop a list of all related articles to our research questions and identify appropriate category. Sentiment analysis has many strategies and techniques were used, such as machine learning, polarity lexicons, natural language processing, and psychometric scales, which determine different types of sentiment analysis, such as assumptions made, method reveals, and validation datasets. This component is extremely helpful for finding the similar word and word-analogy. It is extended to topic classification and detection by combining all word vectors within the sentence or paragraph.

Over the last decade, a lot of data has been accumulating from social media. It is vital point of data mining. Twitter is more suitable for data extraction than other social media, they give us API to extract data. Twitter posts are basically within 140 characters and different types of company and people use. It is little bit different others social media. We collected data Bangladeshi top company Grameenphone and Robi hashtags data and research with them. Our focus is basically ranging of corpus which use twitter post. It is a little difficult to find business word among them because they use a lot of words which do not match business vocabulary.

To overcome all the problems, we reframe our business trend model which gives us business idea. Then we research how much positive or negative and our result. Finally, we can show our final result and their positive or negative percentage.

2.2 Related Works

Sentiment analysis has been studied in a wide range of area such as politics, finance, business, media, and others. It is very useful to movie review, teaching review, hotel review. This paper creates a system architecture to enhance the accuracy of the sentiment cost, the feedback passed through polarity identification, negation tagging, taking into account the phrases surrounding a conferred word [1].Aung, Khin Zezawar, and Nyein Nyein Myo have developed emotion vocabulary or sentiment lexicons is using for opinion mining and sentiment analysis [2].

This vocabulary is used as data source how similar our processing data for machine learning approach. Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun has developed text sentiment analysis is an automated system to figuring out whether a textual content segment carries objective or opinionated content, and it is able to furthermore decide the text's sentiment polarity [3].Twitter, Facebook has most of relevant topic to analyze data. In this paper [4] Prakruthi, V., D. Sindhu, and S. Anupama Kumar discusses about the real-time twitter analysis by tokenization.

Tokenization divide a posts phrases, words, keywords, symbols, other elements and remove unnecessary data. Xdgd Akter, Sanjida, and Muhammad Tareq Aziz discuss about Facebook get right of entry to token is used to construct a URL to the Facebook Graph API from which Facebook Comments are accrued [5].

The correlation of posts and identify consumer. Lokmanyathilak Govindan Sankar Selvan and Teng-Sheng Moh have developed the framework in paper which uses actual online Twitter facts stream, which might be wiped clear and analyzed after which speedy feedback is received via opinion mining [6].

Banić, Lada, Ana Mihanović, and Marko Brakus discover a method that generated information from big data is intended for the possible prospective customer [7]. People can get decide to produce suitable product. Iqbal, Farkhund develop Genetic Algorithm, design, and compare a hybrid sentiment evaluation framework via combining ML and lexicon-based totally processes so as to solve the constraints of every approach [8].

Machine learning algorithms may be classified into 3 classes referred to as Supervised, Unsupervised and semi-supervised. Supervised getting to know is inductive in nature imply training data consists of the required output. Supervised gaining knowledge of is used in class and regression. In unsupervised learning does not need training records. Unsupervised studying is utilized in clustering and sample popularity [9].

Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun discuss about word attaching acquired with the aid of unsupervised learning on big twitter corpora that uses dormant contextual semantic relationships and co-prevalence statistical characteristics among words in tweets [10]. Harakawa, Ryosuke, Takahiro Ogawa, and Miki Haseyama developed a video algorithm which determine web video groups with same things at multiple abstraction levels [11].

Kumar, S. M., and Meena Belwal discusses as a consequence of the experiments we can understand sentiment awareness. After collecting the data create a tool which performance dashboard showcases the information by understating the business behavior for an organization [12].Poria, Soujanya describes text classification, multimodal emotion recognition and sentiment analysis [13].

Text category is a crucial position in many NLP programs, along with spam filtering, data retrieval, net seek, and ranking and report class. This paper particularly approximately the sentimental analysis of tweets using R language that is helpful for sentimental records within the form of both tremendous rating, poor score or impartial in among them. They execute the analysis of tweets which might be in length of TBs which means that big data the use of the R language and R hadoop Connector [14].

Till now, a brilliant deal of work has been done on vicinity-based investigation utilising word library. For instance, Cruz Laura and his institution (2017) composed a section about making use of lexical library for constant domain [15].

It was created to discover the first-class of assumptions communicated in social web texts. In his paper, he depicted how SentiStrength works making use of lexical technique and the use of its own ideas and phrases. Scientist Soumi Dutta and his organization (2015) labored on sentiment evaluation of on line content the usage of WordNet [16].

Developing countries' small and medium startup are face some problems and Challenges of getting access to clever statistics for choice making at unique kind. Tope Samuel Adeyelure, Billy Mathias Kalema and Kelvin Joseph Bwalya build Mobile Business Intelligence (MBI) systems that textual, structural equation modelling. It helpful for decision making and reduce cost [17].

Amadou Sienou, Achim P. Karduck, Elyes Lamine and Hervé Pingaud build a risk indicator which leads data [18]. Mohammed H. Abd El-Jawad, Raina Hodhod and Yasser M. K. Omar apply sentiment analysis different social media data and run multiple machine learning algorithms [19].

2.3 Research Summary

There are many projects have been done about sentiment analysis and business intelligence. We solve their connected a small problem. It is different from others thesis project. They are use a lot of technology to solve problems but all research target extract emotions. Because emotions are different types of categories.

After studying the paper, we can see a lot of work is done emotion vocabulary mining or sentiment corpus mining [2]. Most of the work done are different types of identification or popularity, build model. They identify consumer from real time twitter analysis and give feedback [6]. Our main focus is business sentiment relevant analysis we can see create data collector tool that collect data to improve showcases data and help to take decision or understanding business process [12].

Day by day data storage accumulated massive amount data and problem is their performance. They build a techniques R language and hadoop increase the performance [14]. It divides sentences tokenization with phrases, words, keywords, symbols, other elements and remove unnecessary data then it estimates sentiment their opinion [5]. In this paper compare several types of machine learning algorithms with sentiment analysis. The hybrid model gives best accuracy [19].

We experiment Bangladeshi top company Grameenphone hashtag data that is one kind of text mining and natural language processing. We collected data Grameenphone, Robi, Daraz with twitter API and find out most use words which are related to business sentiment. We proposed to use machine learning algorithms and natural language classifier.

2.4 Scope of the problem

It is very difficult to expound the sentiment idea from natural language processing including machine learning algorithm. Because social media data has huge opinion spam. Once sentiment analysis beneficial properties popularity as a metric to gauge performance and manufacturer photograph of a company, such mal-practices may end up very common which will lead to diminished reputation of Sentiment Analysis. Lack of information or incomplete information create conflict find out exact sentiment opinion. Sometime it gives result wrong, positive results shows as negative. For example, the blog is anything but useful. It shows negative as a negative sentence. Sometimes data are ambiguous that algorithm cannot understand it is negative or positive. As a human being there are no problem detecting a text positive or negative but machine learning approach has some problems. For that reason, we clean data such as unnecessary, hashtag, URLs and web links, numbers, white spaces, uppercase characters delete from our dataset. However, dataset has some problem sarcasm, ambiguity, etc. but it is not a big fact for sentiment analysis. Then surveillance our main focus discovers business trend. It is more helpful business man whom grow up good business.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This research paper we analysis numerous company data and identify their trade tendency in recent time. We can also identify how much tweet positive negative and neutral. We use data mining techniques to predict their propensity in this we will discuss our methodology how we collect data and analysis. We also discuss what is our implementation issues and requirement. This part we demonstrate our model and procedure step by step. Different kinds of technologies are used for tweet sentiment analysis.

3.2 Research Subject and Instrumentation

Our research subject is integrating sentiment analysis and business intelligence with natural language processing to find out business trend. Although many projects have been done about business intelligence and sentiment analysis but we integrate two methods do a small portion solve. It called that every tiny problem solves bring huge success. This process can be used in other languages and other business company. We use it only Bangladeshi company and English language. For that reason, we picked different Bangladeshi company data from twitter by twitter API Which provide us their company hashtag data.

Collected data analysis with Jupyter notebook and Spyder IDE. We apply for Twitter Api in twitter developer site. Then they give us Api and we collect company name hashtag data. That’s file is csv format. We research on that data is various algorithms and natural language classifier such as Naïve Bayes, Bag of Words, Bigrams, Word2vec.These technologies are used for better solution.

3.2.1 Data Cleaning and Pre-Processing

The noise data create conflict to detect sentiment and data are mixed each other tweet. Twitter data contains huge amount of link, URL, hashtag, new line, punctuation unnecessary words. We take aside that kind of data and retain only vocabulary and emotion words. Some tweets are mixed numerous types language we count English language text.

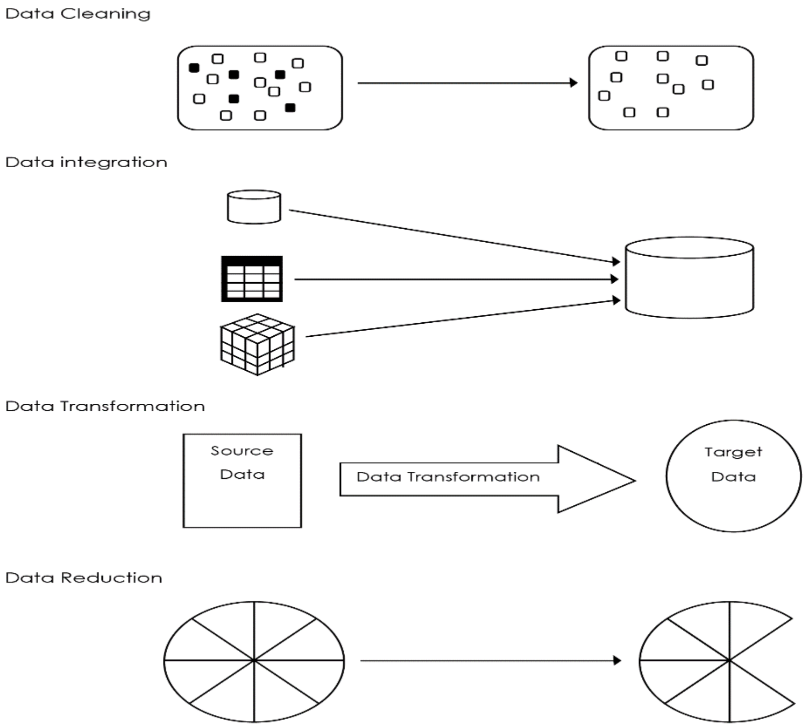


Fig3.2.1: Data Preprocessing Technique

Data Cleaning is the first step of preprocessing. Techniques to clean link's list to evict unvalued links are of importance for any type of link analysis. By checking the type of the "href" html tag eviction of unvalued links can be reasonably accomplished. For instance, Table 1 shows the all links with type like, gif, jpeg, GIF, JPEG, jpg, JPG, and map can be evicted. So, in structure mining data cleaning means selection of links related to web pages and active.

Extricated links, from various server and their pertinent data required for process put away together. Transform the gathered information in a novel arrangement for process ahead. Data decrease implies chooses just that data required for calculation among loads of data accessible on website page.

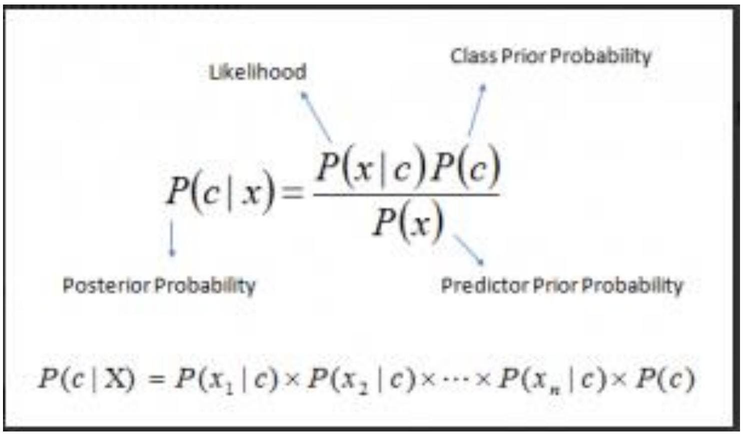
BEFORE CLEANING	AFTER CLEANING
@TelenorGroup: Connectivity Matters. Find out how Telenor and the telecommunications industry is impacting productivity, economic activi\xe2\x80\xa6'.	connectivity matter finds Telenor telecommunication industry impacting productivity economic activi
b'RT @Michael_Telenor: We are happy to host our industry friends who believe the fair representation of men &amp; women at work can ensure produc\xe2\x80\xa6'.	happy host industry friend believes fair representation men amp woman work ensure produce
b'RT @Michael_Telenor: Good news from BD. I am grateful to the Ministers of Finance, of Posts &amp; Telecoms, the Chairs of the National Board o\xe2\x80\xa6'.	grateful minister finance post amp telecom chair national board
@Grameenphone\xe2\x81\xa9. Passionate and committed to take Digital Ban\xe2\x80\xa6'.	passionate committed take digital ban
b'Ookla @Speedtest recognizes #Grameenphone the fastest mobile network of Bangladesh @Michael_Telenor\xe2\x80\xa6 https://t.co/aRE4oRuh3T'	ookla speed test recognizes Grameenphone fastest mobile network Bangladesh Michael Telenor
b'RT @TelenorGroup: Axiata and Telenor have agreed to end discussions regarding a non-cash combination of their telecom and infrastructure as\xe2\x80\xa6'.	Telenor group axiata Telenor agreed end discussion regarding non cash combination telecom infrastructure
b'RT @TelenorGroup: NEWS JUST IN FROM BANGKOK: \nInnovative mobile app, AgriMatch, wins the Telenor Youth Forum and USD 15,000 seed funding.\xe2\x80\xa6'.	Telenor group news Bangkok innovative mobile app Agri match win Telenor youth forum usd seed funding

Table 3.2.1.1: Data Cleaning and Pre-Processing

The table presents data cleaning and pre-processing. This table shows about 30% to 40% data remove from our text. Thus, we remove 1991 Grameenphone and 1685 Robi tweets and get vocabulary.

3.2.2 Naïve Bayes Algorithm

Naïve Bayes classifiers are based on Bayesian arrangement strategies. These depend on Bayes' hypothesis, which is a condition depicting the relationship of contingent probabilities of measurable amounts. In Bayesian characterization, we're keen on finding the likelihood of a mark given some watched highlights, which we can compose as P(C | X). Bayes' hypothesis discloses to us how to express this as far as amounts we can process all the more legitimately:



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig3.2.2: Naïve Bayes Formula

Here,
P(Sentences | Polarity) = P(Polarity | Sentences) * P(Sentences) / P (Polarity)

P(Sentences 1 / class A) = Sentences / ni(A)

P(Sentences 1 / class B) = Sentences / ni(B)

P(Sentences 2 / class A) = Sentences / ni(A)

P(Sentences 2 / class B) = Sentences / ni (B)

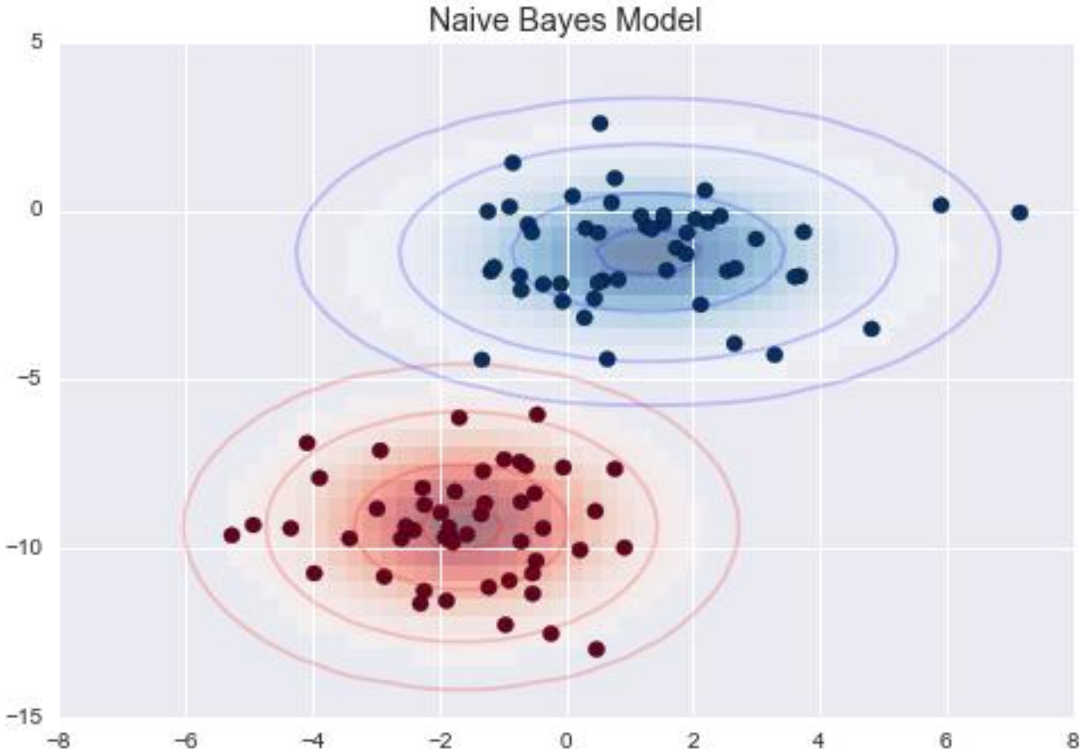


Fig 3.2.3: Naïve Bayes Model

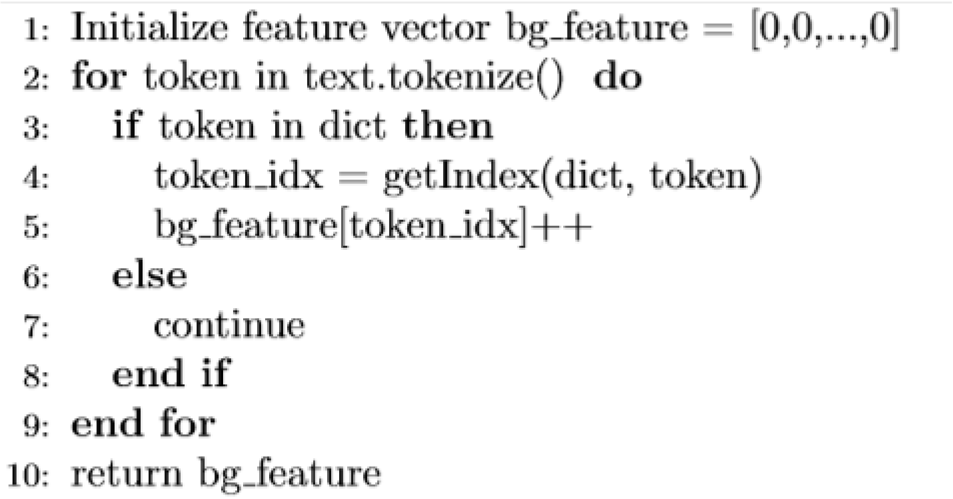
It is utilized as a probabilistic learning strategy for content order. The Naive Bayes classifier is one of the best-known calculations with regards to the grouping of content reports, i.e., regardless of whether a book archive has a place with at least one classifications (classes).

It is a case of content characterization. This has become a prevalent component to recognize spam email from real email. A few current email administrations execute Bayesian spam separating. Numerous server-side email channels, for example, DSPAM, SpamBayes, SpamAssassin etc. utilize this strategy.

It tends to be utilized to dissect the tone of tweets, remarks, and audits—regardless of whether they are negative, positive or impartial. The Naive Bayes calculation in mix with cooperative separating is utilized to fabricate cross breed suggestion frameworks which help in foreseeing if a client might want a given asset or not.

3.2.3 Bag of words

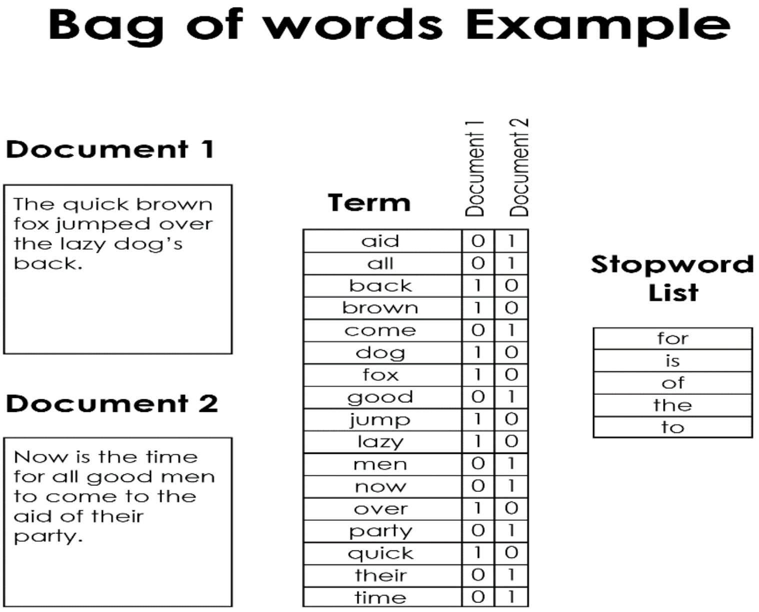
The bag of-words model is a disentangling portrayal utilized in regular language handling and data recovery (IR). In this model, a book, (for example, a sentence or an archive) is spoken to as the pack (multiset) of its words, ignoring syntax and even word request yet keeping assortment.



```
Initialize feature vector bg_feature = [0,0,...,0]
for token in text.tokenize() do
    if token in dict then
        token_idx = getIndex(dict, token)
        bg_feature[token_idx]++
    else
        continue
    end if
end for
return bg_feature
```

Fig3.2.3: Pseudocode of Bag of Words

Bag of word discover frequency of each word from whole dataset. It counts word score current document frequency.



Bag of words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List
for
is
of
the
to

Fig3.2.4: Bag of Words Example

3.2.4 NGram

1.Unigram model ignored context:

$$P(wi|w0\ldots wi-1) \approx P(wi)$$

2.Bigram model adds one word of context:

$$P(wi|w0\ldots wi-1) \approx P(wi|wi-1)$$

3.Trigram model adds two words of context:

$$P(wi|w0\ldots wi-2wi-1)$$

3.2.5 Implementation Procedure

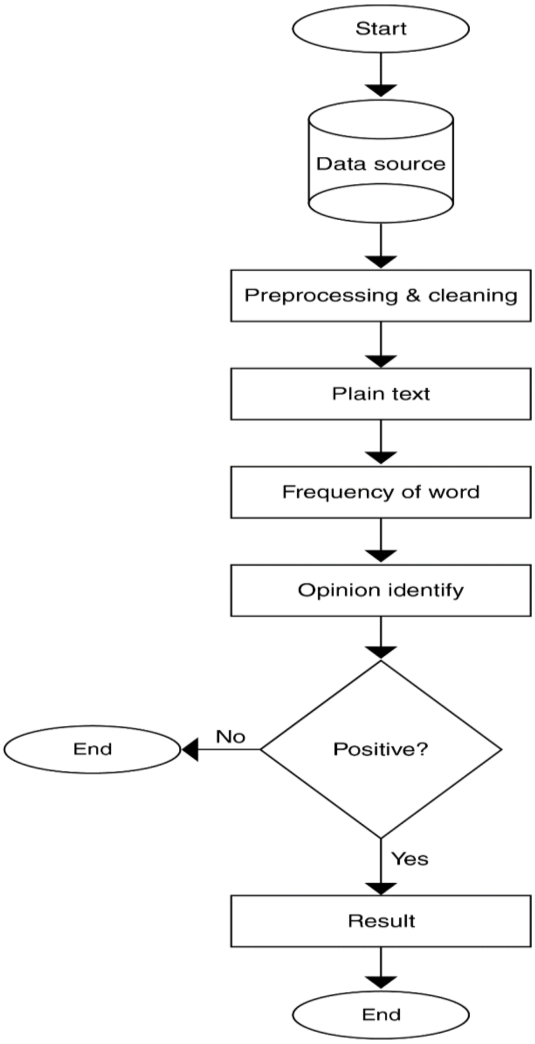


Fig 3.2.5: Implementation procedure flow chart

3.3 Data Collection Procedure

There are many social media sites gives data like Facebook, twitter etc. we select twitter for takeout data. Twitter is very popular for providing Api and easily get data. It is very well for sentiment analysis and create structure, framework with opinion mining. There are many tools to pick up twitter data. We use twitter Api and python script to collect data. First of all, we apply twitter developer site for developer account. They bring a lot of information from us; we open this account student research purpose. Then we create Api for took data. Api has four key such as consumer key, Consumer secret, access key, access secret. This key called access token which can request for data on twitter database.

Python language has different types of library for easier to use various working projects. We use tweepy library for accessing the Twitter API. Tweepy helps to accessing twitter with Basic authentication and the newer method, oauth. Twitter give real time data. We run our script and collect different company hashtag data. Dataset is csv for format and it has three attributes such as id, created date and text. Text means their tweet data. Real time data is very essential for us it describes recent topic which very effective for analysis. We collected 1991 tweets Grameenphone hashtag data. We collected others company data also but we research only Grameenphone and Robi data. This data use for test purpose to generate new trend and give positive or negative output. We collected training data from internet such as movie, product review, google play store comment etc. That data use for train. That dataset each sentence denoted by 0 and 1. Positive sentences determined by 1 and negative sentence determined by 0.

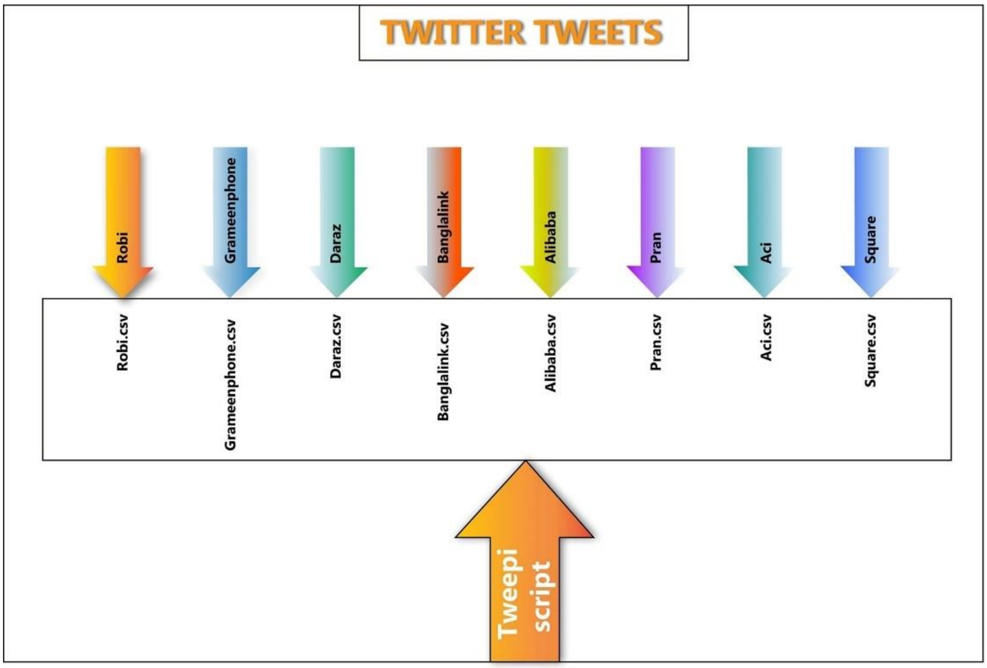


Fig 3.3.1: Twitter Dataset from Api

We collect 8 different companies data set with Twitter Api and python script. This methodology can apply any companies. We apply some dataset find out our goal.

3.4 Statistical Analysis

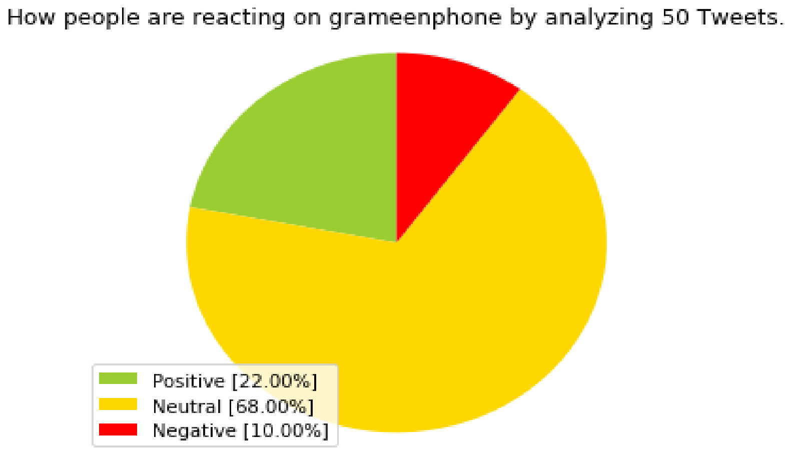
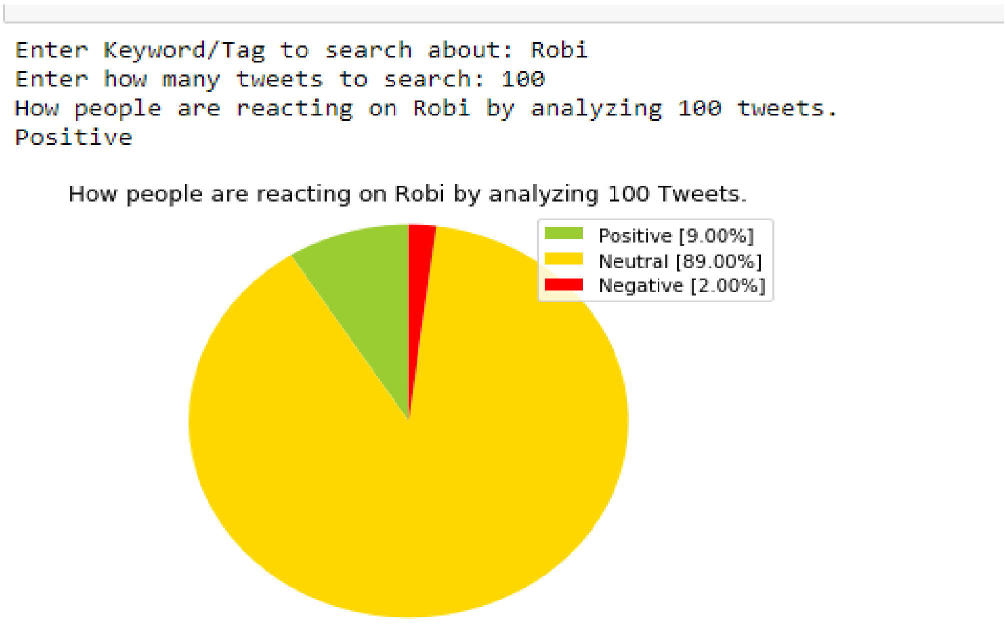


Fig 3.4.1: Grameenphone polarity

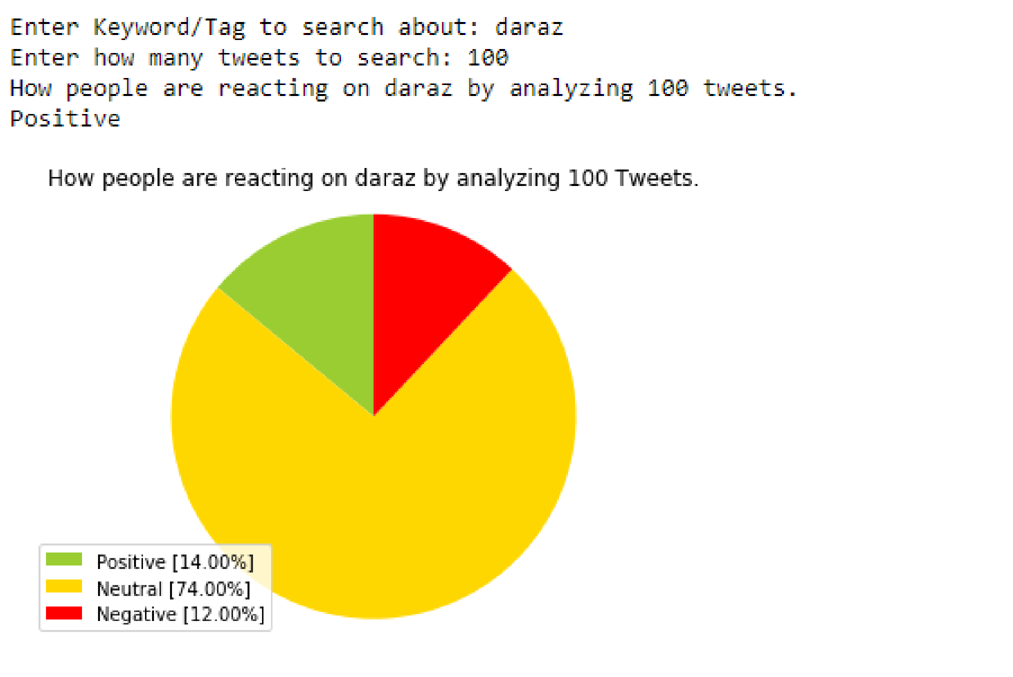
Statistical analysis is significant for any kind of research. It is contributing about the generate idea of the project. We run python script with tweepy, textblob library. Textblob give polarity any sentences. Tweepy is use for easy access twitter database with twitter API. We use matplotlib which make different types of graphs. All of library and package made this figure 3.4.1, 3.4.2 and 3.4.3.



```
Enter Keyword/Tag to search about: Robi
Enter how many tweets to search: 100
How people are reacting on Robi by analyzing 100 tweets.
Positive
```

Fig 3.4.2: Robi Polarity

We can see most of the figure express neutral. Because people cannot their express opinion positive and negative way. Their opinion is descriptive, narrative, questionnaire that means not determine positive or negative polarity. We can see Grameenphone company positive is 22% and negative 10%. It is a good company, customer satisfied company. But it has some lacking about 10% negative. We survey some areas has networking problem, high price, etc. For that reason, they get some negative impact. Robi has 9% positive and 2% negative that means it is better than Grameenphone. But grameenphone is popular company. They first start sim service in Bangladesh. Other companies try to reach like Grameenphone that's why they give many opportunities and benefits. Daraz is a large online shopping website in Bangladesh. They have 14% positive and negative 12% feedback. Sometimes we listen they give poor product. They have many allegations about product quality and service.



```
Enter Keyword/Tag to search about: daraz
Enter how many tweets to search: 100
How people are reacting on daraz by analyzing 100 tweets.
Positive
```

Fig 3.4.2: Daraz Polarity

The charts provide us general summary about some companies where we can get an idea about them. How much positive and negative impact we can gather knowledge. Next, we generate the idea business trend.

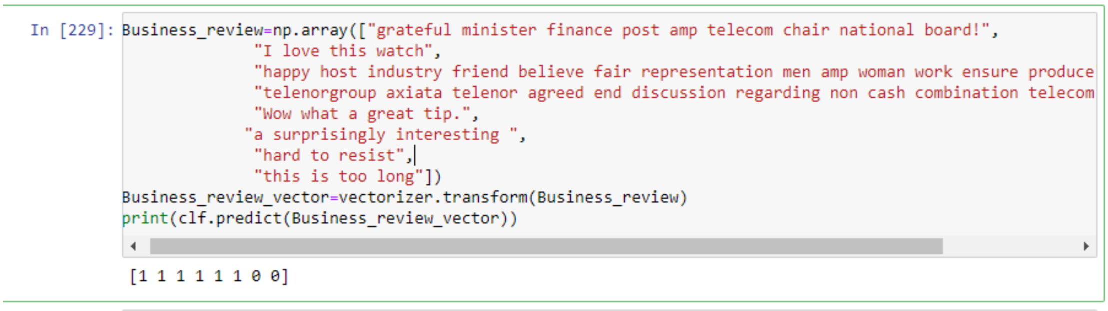
3.5 Implementation Requirements

All the data are csv format and it has three attributes, we copy the text attribute from csv file. Preprocess all the data that means clean unnecessary, URL, hashtag, etc. Data. Then we can measure the frequency of word. We remove it python scripts with import stop words which remove unnecessary words.

We use NumPy, Pandas, Matplotlib, Sklearn etc. package to solve the methodology. All the process has been done 64bit, 4GB ram, corei3, windows 10 machine. All the process describes step by step.

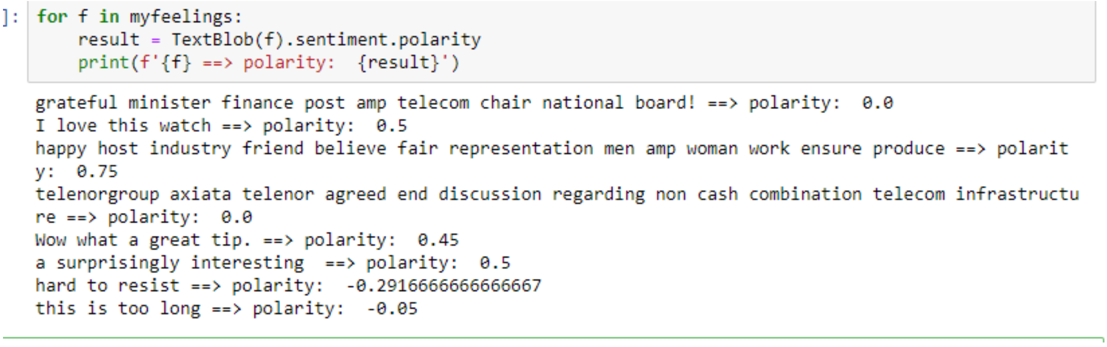
3.5.1 Splitting Training and Test data

After cleaning and pre-processing we test our data. It is not training data, this data train data. Training data is predefined because machine learning algorithm can not understand human opinion. We need to teach them what is positive and negative data. For that reason, we predefined every sentence which positive or negative. We train data and positive, negative classify with Naïve Bayes algorithm. We also determine every sentences polarity that means a sentence how much positive sentence and negative sentence.



```
In [229]: Business_review=np.array(["grateful minister finance post amp telecom chair national board!",
                         "I love this watch",
                         "happy host industry friend believe fair representation men amp woman work ensure produce
                         "telenorgroup axiata telenor agreed end discussion regarding non cash combination telecom
                         "Wow what a great tip.",
                        "a surprisingly interesting ",
                         "hard to resist",
                         "this is too long"])
          Business_review_vector=vectorizer.transform(Business_review)
          print(clf.predict(Business_review_vector))

           [1 1 1 1 1 1 0 0]
```

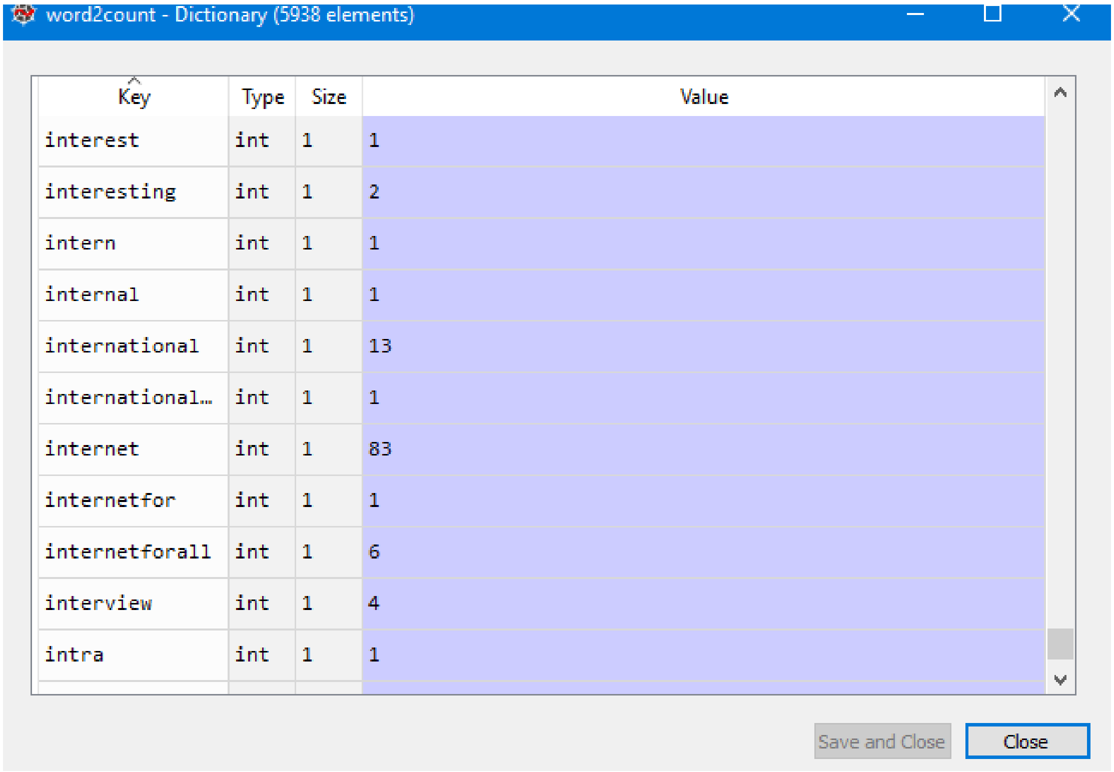
Fig 3.5.2.1: Ratio of Training and Testing



```
]: for f in myfeelings:
       result = TextBlob(f).sentiment.polarity
       print(f'{f} ==> polarity:  {result}')

   grateful minister finance post amp telecom chair national board! ==> polarity:  0.0
   I love this watch ==> polarity:  0.5
   happy host industry friend believe fair representation men amp woman work ensure produce ==> polarit
   y:  0.75
   telenorgroup axiata telenor agreed end discussion regarding non cash combination telecom infrastructu
   re ==> polarity:  0.0
   Wow what a great tip. ==> polarity:  0.45
   a surprisingly interesting  ==> polarity:  0.5
   hard to resist ==> polarity:  -0.29166666666666667
   this is too long ==> polarity:  -0.05
```

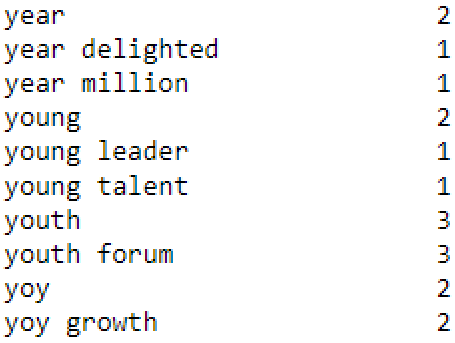
Fig 3.5.2.1: Ratio of Training and Testing

The two figures show how much a sentence positive and negative. They identify their opinion each text. One figure is done by Naïve Bayes algorithm Other is python library Text blob. Text blob is determining how much polarity has a sentence. Many sentence text blob cannot determine. We compare the two figures we can see first sentence is a positive. Text blob give result 0 that means it is a neutral emotion. Because it does not determine their package. We develop our dataset which has this sentence. That's why it gives positive result. Fig 3.5.2.1 it gives result 1 that means it is positive result and we done this result Naïve Bayes algorithm.



Key	Type	Size	Value
interest	int	1	1
interesting	int	1	2
intern	int	1	1
internal	int	1	1
international	int	1	13
international…	int	1	1
internet	int	1	83
internetfor	int	1	1
internetforall	int	1	6
interview	int	1	4
intra	int	1	1

Fig 3.5.2.3: Ratio of Training and Testing



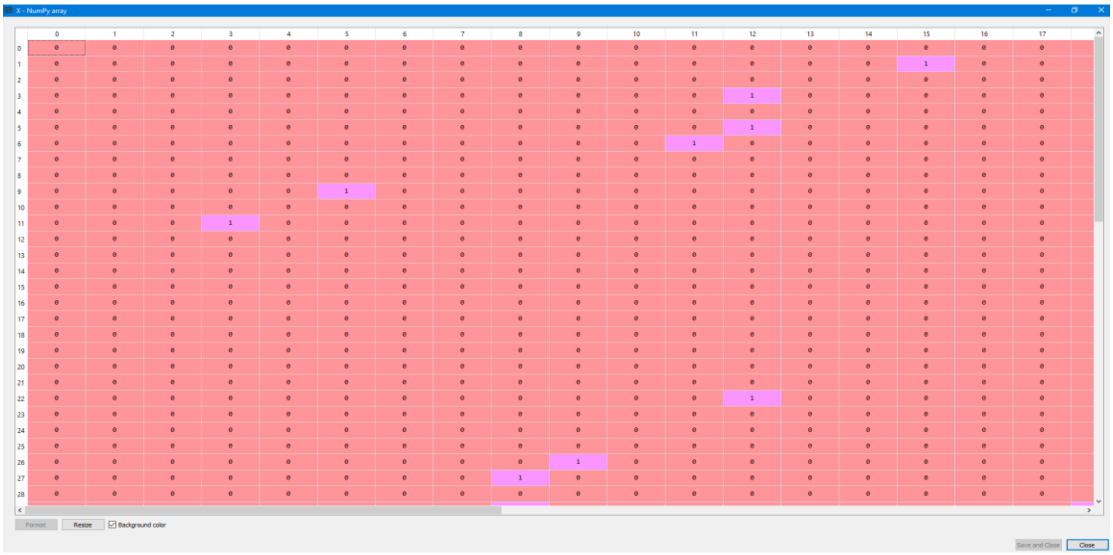
```
year                        2
year delighted              1
year million                1
young                       2
young leader                1
young talent                1
youth                       3
youth forum                 3
yoy                         2
yoy growth                  2
```

Fig 3.5.2.4: Ratio of Training and Testing

We see two picture Fig 3.5.2.3 is text frequency which is done by Bag of words that natural language processing classifier. We run all the preprocessed text which we collect from twitter. That shows frequency of words. Fig 3.5.2.4 shows N-gram classifier that show frequency one word and they're beside word. Bag of word represent multi-set of words. Both algorithms are need for us then we compare that two algorithm and bring well result.

3.5.2 Feature Selection

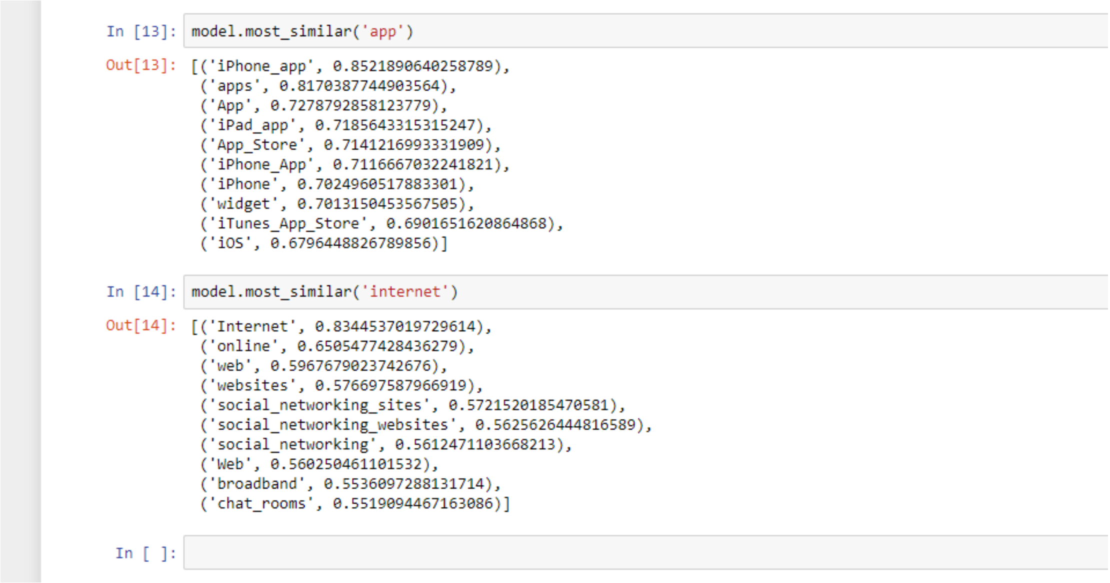
Feature selection is essential for us. Feature start from preprocessing to algorithm run and find out best output. We have passed 3676 Robi, Grameenphone tweet in the bag of word and n-gram classifier. They gave out frequency of words. We used on n-gram count vectorizer from Scikit-learn which makes vector of occurrence count of words. Bag of words use word2count feature divide every sentence is single word. It provides matrix of every unique worlds.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig 3.5.3.1: Sparse matrix

This matrix contains most of the value is 0, that is called sparse matrix and most of the value are non-zero that is called dense matrix. Big sparse matrices are useful in especially in applied machine learning approach, data counts, data encodings that map categories to counts, and even in whole subfields of machine learning such as natural language processing. Bag of get better result than bigram, trigram. It get frequency a unique word that we determine the expected outcome better way.



```
In [13]: model.most_similar('app')

Out[13]: [('iPhone_app', 0.8521890640258789),
          ('apps', 0.8170387744903564),
          ('App', 0.7278792858123779),
          ('iPad_app', 0.7185643315315247),
          ('App_Store', 0.7141216993331909),
          ('iPhone_App', 0.7116667032241821),
          ('iPhone', 0.7024960517883301),
          ('widget', 0.7013150453567505),
          ('iTunes_App_Store', 0.6901651620864868),
          ('iOS', 0.6796448826789856)]

In [14]: model.most_similar('internet')

Out[14]: [('Internet', 0.8344537019729614),
          ('online', 0.6505477428436279),
          ('web', 0.5967679023742676),
          ('websites', 0.576697587966919),
          ('social_networking_sites', 0.5721520185470581),
          ('social_networking_websites', 0.5625626444816589),
          ('social_networking', 0.5612471103668213),
          ('Web', 0.560250461101532),
          ('broadband', 0.5536097288131714),
          ('chat_rooms', 0.5519094467163086)]

In [ ]:
```

Fig 3.5.2.1: word2vec similar word detect

Word2vec use to finding similar word and it is a technique that use for word embedding. It divides each word vector wise. Each word retains one unique vector that find out most similar words mathematically. Word2vec model learns to map each binary word id into a low-dimensional continuous vector-space from their distributional properties observed in some raw text corpus.

Thus, way we find our find out business trend. First of all, preprocess and clean the data for plain text. This data use for find out highest frequency of word. That highest frequency word uses for similar words. Then every word relevant sentence check for positive or negative output. It is way we can get better efficiency.

3.5.3 Algorithms

The research is totally based on experiment and used supervised machine learning. Our dataset is unstructured where collect from twitter. Another is train data collected from internet, product reviews, movie reviews, google paly store data. This dataset train for positive or negative sentence detection. We Naïve Bayes algorithm to get better output because it works with how much probability has each sentence and give better feedback to learn output.

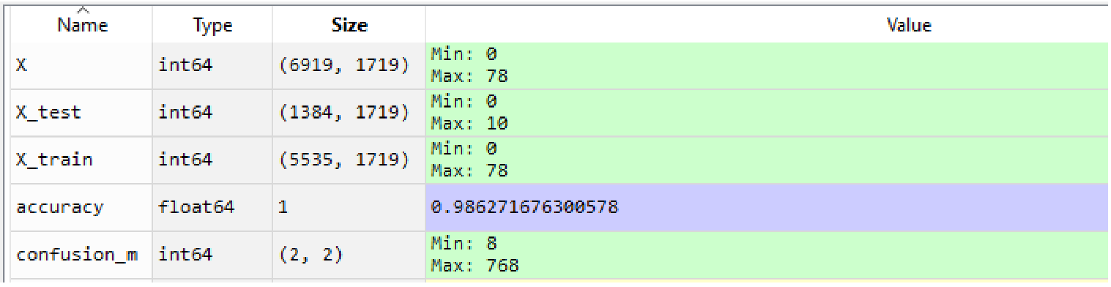
CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This studying sentiment and business intelligence with machine learning approaches. We apply machine learning algorithm and natural language processing classifier such as Bag of words, N-gram, word2vec. We apply text blob for get polarity each sentence. All of things bring well solution. Because each technology is expert in a particular subject.

4.2 Experimental Results

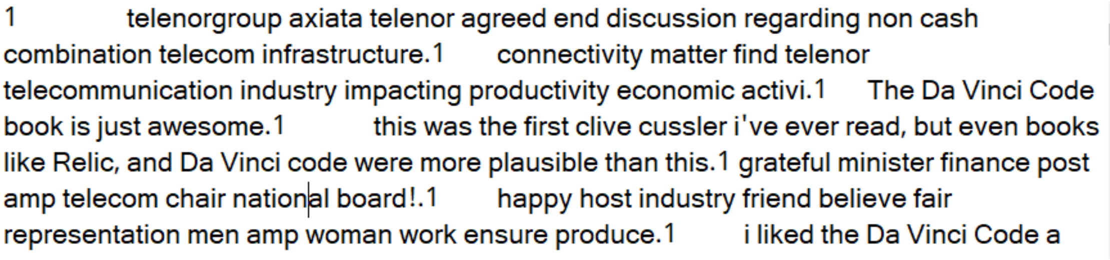
This experiment uses Naïve Bayes algorithm for train data train and testing. We use latest dataset format. This always give good result for text classification. It predicts the probabilities for each class each sentence that given information or record belongs to particular subject or class.



Name	Type	Size	Value
X	int64	(6919, 1719)	Min: 0 Max: 78
X_test	int64	(1384, 1719)	Min: 0 Max: 10
X_train	int64	(5535, 1719)	Min: 0 Max: 78
accuracy	float64	1	0.986271676300578
confusion_m	int64	(2, 2)	Min: 8 Max: 768

Fig 4.2.1: Accuracy with Naïve Bayes

This figure shows accuracy of our train dataset. It gets 98% accuracy for our data set. It is well work for our test sentence. We use best algorithm for text classification and best dataset format that’s why it gives best output.

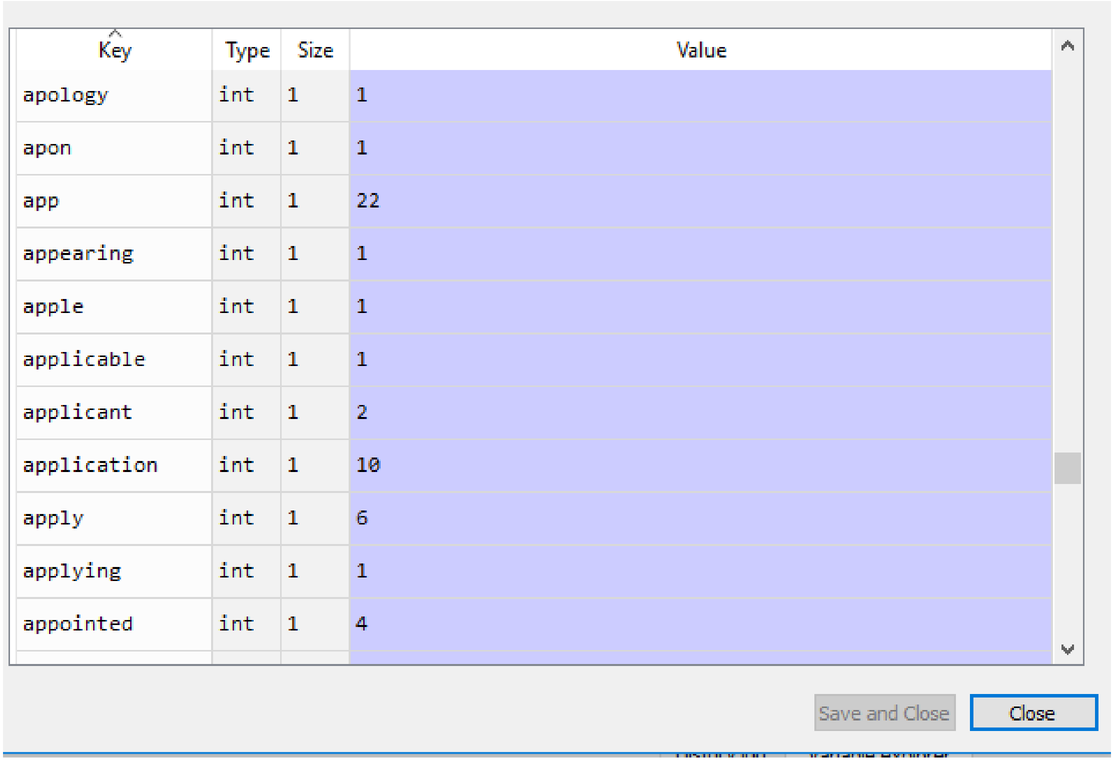


```
1	telenorgroup axiata telenor agreed end discussion regarding non cash
combination telecom infrastructure.1	connectivity matter find telenor
telecommunication industry impacting productivity economic activi.1	The Da Vinci Code
book is just awesome.1	this was the first clive cussler i've ever read, but even books
like Relic, and Da Vinci code were more plausible than this.1 grateful minister finance post
amp telecom chair national board!.1	happy host industry friend believe fair
representation men amp woman work ensure produce.1	i liked the Da Vinci Code a
```

Fig 4.2.1: Dataset format

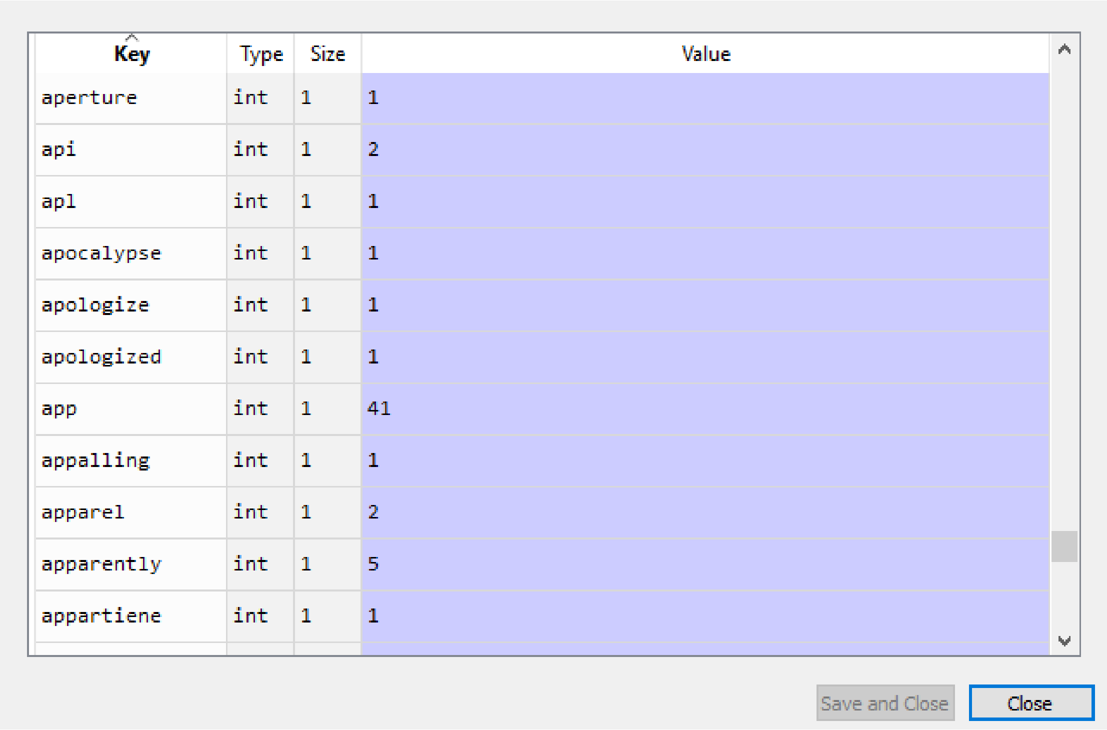
The dataset format positive sentence denoted by 1 and negative sentence is 0. Dataset extension is txt which easy to load any project. There are about 300 plus sample sentences.

4.3 Descriptive analysis



Key	Type	Size	Value
apology	int	1	1
apon	int	1	1
app	int	1	22
appearing	int	1	1
apple	int	1	1
applicable	int	1	1
applicant	int	1	2
application	int	1	10
apply	int	1	6
applying	int	1	1
appointed	int	1	4

Fig 4.3.1: Grameenphone tweet frequency



Key	Type	Size	Value
aperture	int	1	1
api	int	1	2
apl	int	1	1
apocalypse	int	1	1
apologize	int	1	1
apologized	int	1	1
app	int	1	41
appalling	int	1	1
apparel	int	1	2
apparently	int	1	5
appartiene	int	1	1

Fig 4.3.2: Robi tweet frequency

We can see two figure shows Grameenphone and Robi mobile operator company tweet frequency. Fig 4.3.1 shows highest frequency word is ‘app’ that used 22 times and another is ‘application’. That means the recent time they emphasize app. We can watch TV and social media many advertisements with their app such as gp app, wow box. Fig 4.3.2 shows Robi mobile operator company tweet frequency. That is shows ‘app’ word is use 41 times. That means their intension is work with app. They want to people can addict as their own app. In Fig 3.5.2.3 we show most frequency word is ‘internet’. It also related to the app. However, their business is internet. Now we can say online many companies addicted to online business.

Recent time a lot of company build own app. They try to grow up own business with online app. Bangladeshi many advertisements are based on mobile app such as Bkash, Uber, Pathao, Food panda etc. It trends already started when anyone startup good business they build own app to impress, catch and service their customer.

4.3.1 Trend

This experimental research shows business trend for Bangladeshi company. Our topic discovers world business trend. This method useful for find out anyone interested area business data. We show two company trends of their recent time. They emphasize their business with app.

4.4 Summary

There are used about one machine learning algorithm and three natural language processing classifiers to get a well solution. We also use python library text blob that give polarity each sentence. The most important is Bag of word which give that word matrix and word frequency. We research how much positive and negative sentences. Most of the machine package import which helpful to find out get solution easily.

CHAPTER 5
SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the Study

In this research we endeavour to trace positive, negative feedback from Social Media's data exploit sentiment analysis and also trying to discovery new business trend. Studying several tweets, online data and categorize sentence which word mean business related speech and train that sentence for find out positive and negative sentence.

Popular company day by day switch or update their service. Because people like new and unique things. Company try to bring variations their products. A company does not stay similar function at all time, it changes assessment quality of product. So, their need discover popularity of product at a time period and real time data is that obvious time. We used Twitter API using Push method for collected data. Using this API, we taken real time csv format tweets of a company hashtag.

People use different types of languages such as France, Bangla, German, Spanish etc. language. We collect hashtag data that means other languages sentence involve this project. We take only English languages data and do ignore other language data. It was hard to us collect large amount of data. We have got approximately thirty hundred of tweets and online data for our work.

We apply machine learning algorithm naïve Bayes and natural language classifier bag of word, N-gram. Naïve Bayes algorithm classify the sentence positive and negative. Bag of word find out frequency of word that we perceive theme of business.

Finally, our mentioned thesis project is successful and we get 98% accuracy.

5.2 Conclusions

This study, however the way that in a very limited timeframe, has made the issue completely clear. We dissipated on making the issue scope clear and it's fills as a phase for essential increment to this framework. The projects of the capacities of the understudies will help the authority of have a strong diagram of the understudies. The studies are also expected to assure appropriate route and academic guides for the understudies who're on a negative knowledge level. Last final results of the studies are emanated by imposing specific, statistical techniques, calculations and algorithms. Who had attention to their underlying stages of programming have sparkled in pretty a whole lot each different vicinity? Practicing of middle programming reasons a extraordinary deal to continue in different technical zones. Also, specialized data with relational abilities activates a balanced career.

5.3 Recommendations

Greatness is record-breaking a work being developed; our proposed task is exactly at its starting stages. Along these lines, a famous exchange of works can be conceivable to it. For improving the permanency, subservience and proficiency of the investigation, besides ingathering of information is essential. The greater the facts are the extra dependable the effects are. Other than, a permitted set is likewise predicted to decrease the over-fitting of the models. Gradually boom fashions can be used on the information to inquire again.

5.4 Implication for Further Study

Currently, the interest of an information mining master is profoundly esteemed. In view of quality of colossal measure of data in our situation. In this way, this is the reasonable time to work with these sorts of profundity information, so another example is familiar to determine distinctive profundity issues. Wistful deciding is a fundamental piece of Machine Learning. The experimental study which we've carried out reputation identification with a catching result is leaving a stable affect in the back of our work. Still we're working with the gadget and we will maintain on operating on the device moreover for a terrific and more exact system.

REFERENCE

[1] A. Sagum, Jessiree Grace M. de Vera, Prince John S. Lansang, Danica Sharon R. Narciso and Jahren K. Respeto, “Application of Language Modelling in Sentiment Analysis for Faculty Comment Evaluation,” Proceedings of the International Multi Conference of Engineers and Computer Scientists, vol.1, March 2015.

[2] Aung, Khin Zezawar, and Nyein Nyein Myo. "Sentiment analysis of students' comment using lexicon based approach." *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE, 2017.

[3] Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for twitter sentiment analysis." *IEEE Access* 6 (2018): 23253-23260.

[4] Prakruthi, V., D. Sindhu, and S. Anupama Kumar. "Real Time Sentiment Analysis Of Twitter Posts." *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*. IEEE, 2018.5. Weaver, Jesse, and Paul Tarjan. "Facebook Linked Data via the Graph API. "Semantic Web 4.3 (2013): 245-250.

[5] Xdgd Akter, Sanjida, and Muhammad Tareq Aziz. "Sentiment analysis on facebook group using lexicon based approach." *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*. IEEE, 2016.

[6] Lokmanyathilak Govindan Sankar Selvan, Teng-Sheng Moh “A Framework for Fast-Feedback Opinion Mining on Twitter Data Streams”, Collaboration Technologies and Systems (CTS), 2015 International Conference, 1-5 June 2015, Page: 314 – 318

[7] Banić, Lada, Ana Mihanović, and Marko Brakus. "Using big data and sentiment analysis in product evaluation." *2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2013.

[8] Iqbal, Farkhund, et al. "A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction." *IEEE Access* 7 (2019): 14637-14652.

[9] Chaturvedi, Saumya, Vimal Mishra, and Nitin Mishra. "Sentiment analysis using machine learning for business intelligence." *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. IEEE, 2017.

[10] Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for twitter sentiment analysis." *IEEE Access* 6 (2018): 23253-23260.

[11] Harakawa, Ryosuke, Takahiro Ogawa, and Miki Haseyama. "Extracting hierarchical structure of web video groups based on sentiment-aware signed network analysis." *IEEE Access* 5 (2017): 16963-16973.

[12] Kumar, S. M., and Meena Belwal. "Performance dashboard: Cutting-edge business intelligence and data visualization." *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*. IEEE, 2017.

[13] Poria, Soujanya, et al. "Convolutional MKL based multimodal emotion recognition and sentiment analysis." 2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016.

[14] Kumar, Sunny, Paramjeet Singh, and Shaveta Rani. "Sentimental analysis of social media using R language and Hadoop: Rhadoop." *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2016.

[15] Laura, C., José, O., Mathieu R., Pascal, P.: Dictionary-based sentiment analysis applied to a specific domain. In: Information Management and Big Data, pp. 57–68, Springer International Publishing (2017)

[16] Dutta, S., Roy, M., Das, A.K., Ghosh, S.: Sentiment detection in online content: a WordNetbased approach. In: Panigrahi, B., Suganthan, P., Das, S. (eds.) Swarm, Evolutionary, andMemetic Computing. SEMCCO 2014. Lecture Notes in Computer Science, vol. 8947.Springer, Cham (2015)

[17] Adeyelure, Tope Samuel, Billy Mathias Kalema, and Kelvin Joseph Bwalya. "Development of Mobile Business Intelligence framework for small and medium enterprises in developing countries: Case study of South Africa and Nigeria." *2016 4th International Symposium on Computational and Business Intelligence (ISCBI)*. IEEE, 2016.

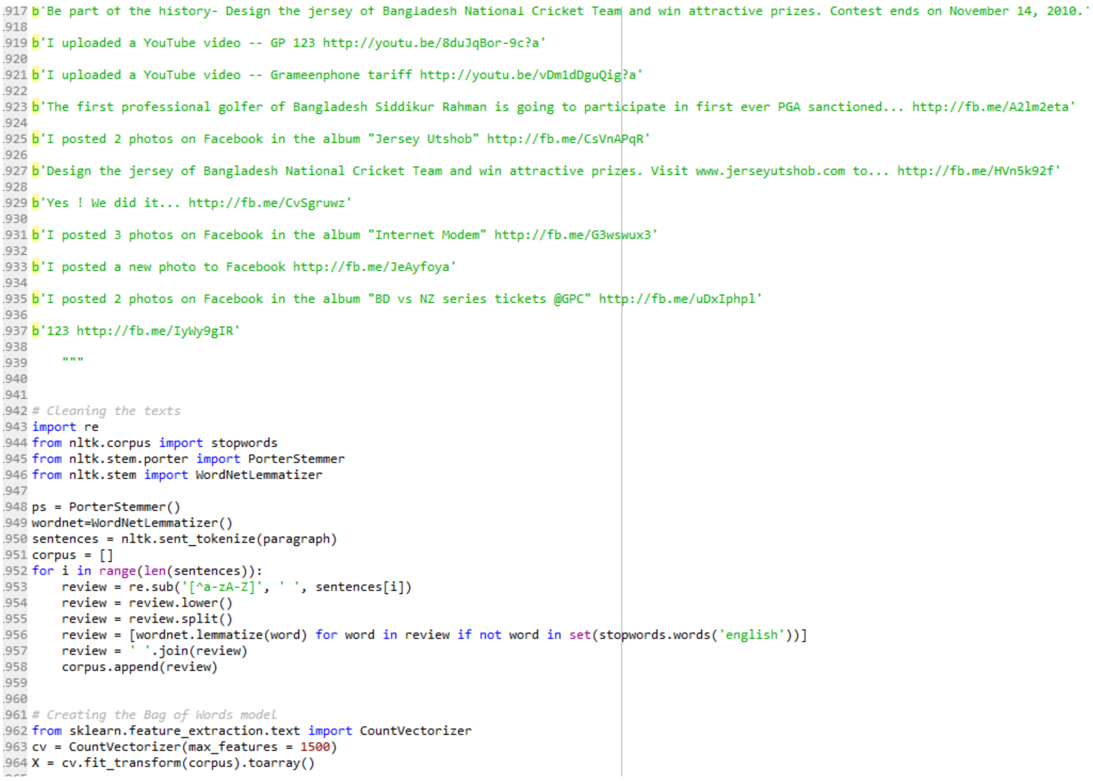
[18] Sienou, Amadou, et al. "Business process and risk models enrichment: Considerations for business intelligence." *2008 IEEE International Conference on e-Business Engineering*. IEEE, 2008.

[19] El-Jawad, Mohammed H. Abd, Rania Hodhod, and Yasser MK Omar. "Sentiment Analysis of Social Media Networks Using Machine Learning." *2018 14th International Computer Engineering Conference (ICENCO)*. IEEE, 2018.

[20] Singh, Kamaljot, Ranjeet Kaur Sandhu, and Dinesh Kumar. "Comment volume prediction using neural networks and decision trees." *IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015), Cambridge, United Kingdom*. 2015.

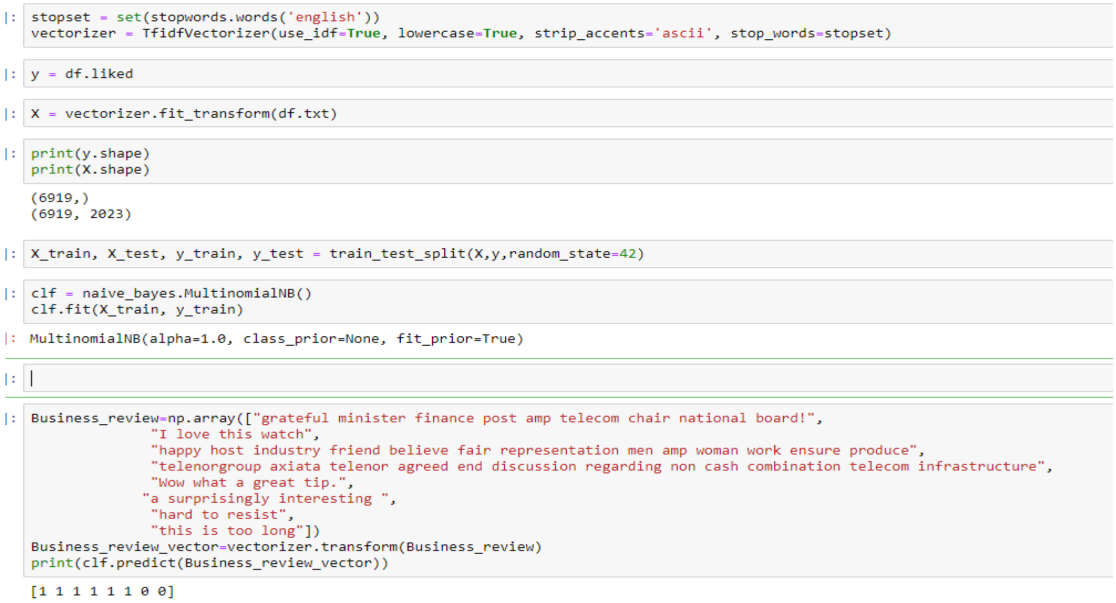
APPENDICES

Appendix A: Machine Learning approach how our method is work.



```
b'Be part of the history- Design the jersey of Bangladesh National Cricket Team and win attractive prizes. Contest ends on November 14, 2010.'

b'I uploaded a YouTube video -- GP 123 http://youtu.be/8duJqBor-9c?a'

b'I uploaded a YouTube video -- Grameenphone tariff http://youtu.be/vDm1dDguQig?a'

b'The first professional golfer of Bangladesh Siddikur Rahman is going to participate in first ever PGA sanctioned... http://fb.me/A2lm2eta'

b'I posted 2 photos on Facebook in the album "Jersey Utshob" http://fb.me/CsVnAPqR'

b'Design the jersey of Bangladesh National Cricket Team and win attractive prizes. Visit www.jerseyutshob.com to... http://fb.me/HVn5k92f'

b'Yes ! We did it... http://fb.me/CvSgruwz'

b'I posted 3 photos on Facebook in the album "Internet Modem" http://fb.me/G3wswux3'

b'I posted a new photo to Facebook http://fb.me/JeAyfoya'

b'I posted 2 photos on Facebook in the album "BD vs NZ series tickets @GPC" http://fb.me/uDxIphpl'

b'123 http://fb.me/IyWy9gIR'

    """


# Cleaning the texts
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer

ps = PorterStemmer()
wordnet=WordNetLemmatizer()
sentences = nltk.sent_tokenize(paragraph)
corpus = []
for i in range(len(sentences)):
    review = re.sub('[^a-zA-Z]', ' ', sentences[i])
    review = review.lower()
    review = review.split()
    review = [wordnet.lemmatize(word) for word in review if not word in set(stopwords.words('english'))]
    review = ' '.join(review)
    corpus.append(review)


# Creating the Bag of Words model
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features = 1500)
X = cv.fit_transform(corpus).toarray()
```

Fig 5.5.1: Bag of words code



```
stopset = set(stopwords.words('english'))
vectorizer = TfidfVectorizer(use_idf=True, lowercase=True, strip_accents='ascii', stop_words=stopset)

y = df.liked

X = vectorizer.fit_transform(df.txt)

print(y.shape)
print(X.shape)

(6919,)
(6919, 2023)

X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=42)

clf = naive_bayes.MultinomialNB()
clf.fit(X_train, y_train)

MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)

Business_review=np.array(["grateful minister finance post amp telecom chair national board!",
                "I love this watch",
                "happy host industry friend believe fair representation men amp woman work ensure produce",
                "telenorgroup axiata telenor agreed end discussion regarding non cash combination telecom infrastructure",
                "Wow what a great tip.",
               "a surprisingly interesting ",
                "hard to resist",
                "this is too long"])
Business_review_vector=vectorizer.transform(Business_review)
print(clf.predict(Business_review_vector))

[1 1 1 1 1 1 0 0]
```

Fig 5.5.1: Naïve Bayes code