BRAND EVALUATION USING SOCIAL MEDIA DATA

by

NAVEEN DHAYANIDHI	**2012103040**
SUNDHARAMURTHY SELLAMUTHU	**2012103075**
SURAJ DUGGIRALA	**2012103076**

A project report submitted to the

FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING

in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ANNA UNIVERSITY, CHENNAI – 25

MAY 2016

BONAFIDE CERTIFICATE

Certified that this project report titled **BRAND EVALUATION USING SOCIAL MEDIA DATA** is the *bonafide* work of **NAVEEN DHAYANIDHI (2012103040)**, **SUNDHARAMURTHY SELLAMUTHU (2012103075)** and **SURAJ DUGGIRALA (2012103076)** who carried out the project work under my supervision, for the fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science and Engineering. Certified further that to the best of my knowledge, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on these or any other candidates.

Place: Chennai
Date:

Dr. S. Valli
Professor
Department of Computer Science and Engineering
Anna University, Chennai – 25

COUNTERSIGNED

Head of the Department,
Department of Computer Science and Engineering,
Anna University Chennai,
Chennai – 600025

ACKNOWLEDGEMENT

We would like to extend our immense gratitude to our project guide, Dr. S. Valli for her perpetual support and able guidance which was instrumental in taking the project to its successful conclusion. We would like to thank the entire committee consisting of Dr. V. Vetriselvi, P. Velvizhy , S.Renugadevi and V. Suganya for constantly motivating us and identifying the areas of improvement on the project. Finally, we would like to thank the Head of the department, Dr. D. Manjula for providing a conducive environment and amenities to facilitate our project work.

Naveen Dhayanidhi Sundharamurthy Sellamuthu Suraj Duggirala

ABSTRACT

A plethora of data accretes on social media everyday. For example, Facebook and Twitter report web data from approximately 149 million and 90 million unique U.S. visitors per month, respectively. With the increasing number of users of social media, the entire marketing concept is getting shifted towards social media such as Facebook, Twitter etc.

Due to its vast size and the number of people using such platforms is highly increasing, it becomes difficult for a particular company to keep track of how exactly the marketing is done for the company and how people view their products online. There are a lot of challenges involved in processing social media data as it is vast, noisy, distributed and highly unstructured.

We have come up with an idea where we value the companies on social media using some efficient mining algorithms of the data collected from the social media sites, based on the following factors such as brand awareness, brand exposure, customer engagement and electronic word-of-mouth.

ABSTRACT

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 SCOPE OF THE PROJECT

Conversations on the Internet produce massive amounts of unstructured data. Both social media and person-to-person information-gathering have value, but social media listening is quickly becoming an important customer intelligence tool. There are several ways to use social media to gain insight, including monitoring online customer support forums, using software tools to gather comments from social outlets such as Facebook and Twitter and encouraging customers to suggest new product features and vote on their favorites.

In a large enterprise, social media monitoring tools can mine text for specific keywords on social networking websites and blogs and in discussion forums and other social media. Essentially, monitoring software transposes specific words or phrases in unstructured data into numerical values which are linked to structured data in a database so the data can be analyzed with traditional data mining techniques.

Organizations use social media monitoring to reach out to customers and prospects for information gathering and front-end customer support. It is used to collect and mine data, especially by organizations seeking customer intelligence to determine current industry trends. The process has become easier - yet more tedious - due to free and readily available outlets, like blogs, wikis, news sites, social networking sites, forums, video/photo sharing sites and message boards.

1.2 ABOUT THE PROJECT

Process

The intricate process of monitoring social media starts with building a corpus, a body of text to be analysed. The corpus is built by using spiders and bots to collect relevant data from social media and the wider web [6]. Although the available tools are increasingly powerful, they do have their limitations. In particular not all of the web is accessible (especially large parts of Facebook due to strict privacy issues) and it is often impossible to determine the exact geographic location of someone posting a comment. Before the collected data can be analysed it must be cleaned [5].

The data that originates from the client, from its various agencies, and from bots needs to be removed. For example, if one of the reasons for conducting a project is to monitor the launch of some new campaign, messages originating from the media, public relation, and marketing agencies need to be removed from the corpus. Another element of cleaning is to remove wrong Matches. For example, when looking at things like soft drinks, coke the drink is good, but coke the drug is not. Once the corpus has been created then the analysis takes place and can include one or more of the following: counts, trends, sentiment, and influence.

Count

The simplest type of analysis is to count however typically key words or terms occur, generally extending this to require words especially contexts. In terms of depth, straightforward counts are fairly superficial, though they need been shown to be of interest in some things, as an example many people have reported success in mapping the frequency of politicians being mentioned and their success in elections. Google have shown success in mapping diseases by numbers the terms that people sort into search engines, to spot attainable incidences of flue and more recently dengue fever.

Trends

In many ways trends are just an extension of counting, looking at whether a term is changing into more used or less used, and to check if it will be coupled to different phenomena. Watching whether or not a term is trending on Twitter has become a key element in brand management, and brands look to examine if they will determine trends, in social media, associated with their campaign launches.

Sentiment-Analysis

Sentiment analysis is either the core advantage of social media observation or the snake oil of 2011, counting on who you confer with. The concept behind sentiment analysis is extremely easy, rather than simply count how often a key word or phrase is employed (for example a brand name), sentiment analysis measures what number of times it's aforementioned during a positive, neutral, or negative manner.

The information collected by the industrial systems are generally too large to code by hand, therefore one amongst the subsequent approaches is followed:

1. Machine-controlled techniques, applying a range of approaches and algorithms.
2. Manually coding a sample of the information.
3. Coding some of the information manually to permit software to learn however to code the remainder.

Influence and Identification

One of the key options of the tools used for social media observance is that, in most cases, they're not designed specifically for market research, they are equally, or maybe more specifically, designed for marketing. Furthermore using the established key words and phrases, the systems will find who is saying what, permitting them to be targeted for selling, viral leads, and word of mouth support.

This power to seek out who is saying what, who is paying attention to whom, and who seems to own influence, could be a two-fold challenge for market research. Firstly, this power risks removing the obscurity of the individuals being researched, and second if market researchers aren't ready to be concerned in uses like intervention and response selling they may realize themselves marginalised within the whole space of social media monitoring.

1.3 CHALLENGES IN BRAND EVALUATION SYSTEM

The brand evaluation system has the following challenges:

1. The crawled data from social media sites such as Facebook and Twitter contains data in a non-human readable format. Conversion of the data to a readable format with specific data tags and the collection of this data was time consuming as the size of the dataset was vast.

2. Piping of the respective python processes to the Ruby On Rails (ROR) app was the second challenge.

3. The evaluated scores had to be presented in a cleaner and creative format such as pie charts, bar graphs, heat map and so forth.

1.4 OVERVIEW OF THESIS

Chapter 2 provides the related works of emotion classification and product comparison systems.

Chapter 3 deals with the overall system architecture and the corresponding module descriptions.

Chapter 4 shows the implementation and results of the proposed system.

Chapter 5 deals with conclusion and future works.

CHAPTER 2

RELATED WORK

The base paper deals with mining Twitter data from people around the world and classifying the tweets based on the sentiments like joy, sadness, surprise, anger and love [3]. It deals only with the sentiment of the people based on their tweets. To identify potential emotional tweets, a large vocabulary of emotion terms was compiled from multiple sources, including the Affective Norms for English Words (ANEW) and the Linguistic Inquiry and Word Count (LIWC) [1]. ANEW provides a set of normative emotional ratings for a set of 1034 English words, and LIWC is text analysis software that calculates the degree to which people use different categories of words across a wide array of texts. In addition to the emotional categorisation, it also attempts to classify tweets based on the location and gender of the user [7]. This information can be very useful in implementing our project.

The reference paper involves a similar concept where the reviews from e-commerce websites is collected for various products such as books, electronics, kitchenware etc and trained the model with that data [4]. They have implemented a machine learning algorithm to classify the reviews, rate the product based on the reviews and then tell if the product is recommended to buy or not. It involves the usage of reviews which are almost of the same format in different e-commerce sites. For the past ten years, there have been many attempts at monitoring the data on social media sites and even the internet. The tools that have been developed have been very useful to several organisations and industries.

Organisations use this to maintain and manage their brands reputation on the internet. This is done by managing what shows up on Google by a process called Search Engine Optimisation. Search Engine Optimisation or SEO works on ranking for certain keywords. Using these keywords, Google can rank sites for search results of these keywords. This is done by Googles algorithm. The first version of this Page Ranking Algorithm was written by Larry Page, the founder of Google. That is where the name Page Rank originated from. However, since then there have been several iterations of this algorithm. The algorithm keeps changing as people started misusing it and getting their low quality sites ranked as high as the number one result on Google.

As time went on and social networking became a trend, Google integrated an element of social into its algorithm. Websites and articles that have a high social equity get ranked higher. Social Equity is based on how many followers or likes a website or article has. The most important criteria is how many shares it has gotten from Facebook, Twitter, LinkedIn, Google Plus and other social networking sites. Shares are considered the most important as people only share relevant and useful and engaging content which is what Google is looking for. Also Likes and Followers can now be faked and bots are being used to play the numbers game.

An early example of the power of social media observation was given in 2005 by the CREEN project [2] , the project monitored the output of one hundred thousand blogs for over 3 years and recorded the dual instances of science connected words with fear/anxiety connected words. The project tracked the amount of hits over time, and when peaks were ascertained the researchers reviewed the key phrases that were driving the will increase.

Samples of spikes within the data were Schaivo (relating to the

Terry Schaivo life support case within the US), and stem (relating to research with stem cells). A second, and additional business example, is provided by the hotel chain, Accord. It has many brands. Some of them are Sofitel, Novotel, and motel 6. Synthesio track 4000 specific Accor hotels, along with 8000 competitors, in eight languages, to provide a worldwide dashboard, 40 regional dashboards, and 4000 hotel specific dashboards, each dashboard displaying key competitors, and every dashboard being updated weekly.

The analysis combines the process of open-ended comments in social media, scores from evaluation sites like Tripadvisor and Booking.com, and additional traditional measures. As a result of using the method, Accor report an increase in brand equity, satisfaction and bookings. The system has allowed Accord to quickly determine underperforming hotels and for individual negative comments to be placed and acted on. Examples of firms that offer social media monitoring platforms include Radian6 and NetBase. Although they have completely different approaches to the matter, each Radian6 and NetBase offer their shoppers (market researchers and brand managers) tools that enable them to look at conversations that are happening on social media that are relevant to their brand.

For instance, NetBase provides ”scorecards,” that take a fast look at however well a brand is doing on social media, and a ”workbench,” which gives firms the flexibility to dig a bit deeper into these queries. Radian6, among alternative services, provides variety of various reports that cover everything from sentiment, to influence and to competitive analysis. People complaining to friends concerning their health or waiting times at hospitals is nothing new. However as more prefer to do so on web forums and social networks like Twitter and Facebook, they will be surprised to find out that hospitals and health care professionals are listening in.

Organisations like the Care Quality Commission, the UK's health and social care regulator, and several other NHS hospitals are beginning to crawl the net for clues concerning wherever they have to analyze low standards or direct further resources. Social media monitoring is turning into common within the private sector, as companies listen out for complaints regarding their own services, or those of competitors, to assist poach customers. In the public sector, eavesdropping on on-line conversations will tap opinions from those who might not wish to fill in a formal survey form. A lot of works associated with mining social media information have been done in the recent past.

CHAPTER 3

SYSTEM DESIGN

3.1 PROPOSED SYSTEM

The proposed system will ask the user to give a brand name that requires to be evaluated. Relevant information about the brand will be scraped from social media sites like Facebook and Twitter. Customer and user feelings towards the brand will be analysed from the preprocessed crawled data. The data visualisation aspect of the project will include a heat map of a world that potrays the performance of the brand around the world in terms of user engagement and user reach.

3.2 SYSTEM ARCHITECTURE

This system consists of three phases. The first phase is the Data Collection and Preprocessing phase. In this phase, the user enters the name of a brand that is to be evaluated. Once the name has been entered the crawler runs through Facebook and Twitter to get relevant information regarding that brand. The output file will be JavaScript Object Notation (JSON) formatted. This file has to be converted to a more readable format such as Comma Seperated Values (CSV). The data is then preprocessed to remove inconsistent and dirty data. It also makes it easier to extract important features that aid in the classification procedure.

The second phase is training and testing the classification model. The preprocessed data from the previous phase is subjected to feature extraction. Initial dataset had to be prepared for both text and numeric data. For text data, the Bag of Words model which will be explained in detail in the corresponding module split-up is used for extracting important features that aids in the training and testing of the text classification model. For numeric data, Support Vector Machine (SVM) model is used for classification against data like count of likes, comments, shares and so forth.

The third phase is data visualisation. The data is crawled real time and predicted against the trained classification models (text and numeric) and the combined score for the evaluation is presented in a graphical format such as a heat map. The web app setup is done using Ruby on Rails and Javascript for graphical elements. Figure 3.1 depicts the process flow of the system architecture.

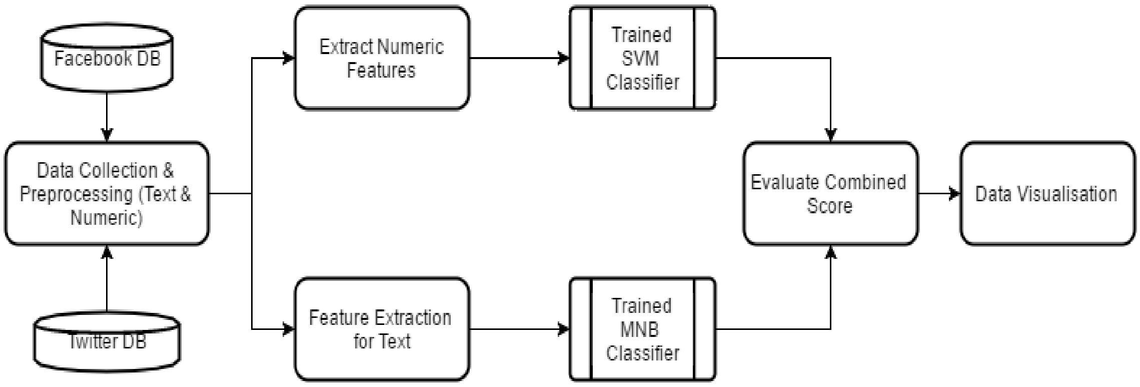


Figure 3.1 System Architecture

1. The crawler module scrapes through the following social media sites:
   - Facebook
   - Twitter

2. Preprocessing module includes:
   - Tokenization
   - Stemming using Porter Stemming

- Lemmatization
- Emoticons processing
- Stop words removal

3. Classification of
   - Text dataset using Multinomial Naive Bayes model
   - Numerical dataset using Support Vector Machine (SVM)

4. Web app built on Ruby on Rails (ROR) to display recommended results.

3.3 LIST OF MODULES

The list of modules are:

1. Data Collection
2. Data Preprocessing
3. Feature Extraction
4. Training and Testing
5. Real Time Crawling & Data Visualisation

3.4 MODULES SPLIT-UP AND DETAILED DESIGN

3.4.1 Data Collection

To collect the data from various social media websites, we have written crawlers in python that carries out this task. Extraction from the social sites is done using the algorithm specified. A tool called the GraphAPI is used to get data out of Facebook. It is a low level HTTP based API that can be used to query data such as post id, posts, comments of posts, user's friends count and likes, shares, comments count of he corresponding post and various other attributes. The output will be JSON formatted. Only relevant data is collected and stored while ignoring redundant and inconsistent data. The relevant data will be stored in CSV format for readability and further preprocessing.

For Twitter, TwitterStreamerAPI similar to Facebook's GraphAPI was used to query relevant data from Twitter. Attributes such as the tweet, follower count, timestamps and so forth are retrieved as JSON and formatted to CSV with the relevant information. Figure 3.2 shows the detailed design for the data collection submodule for text.

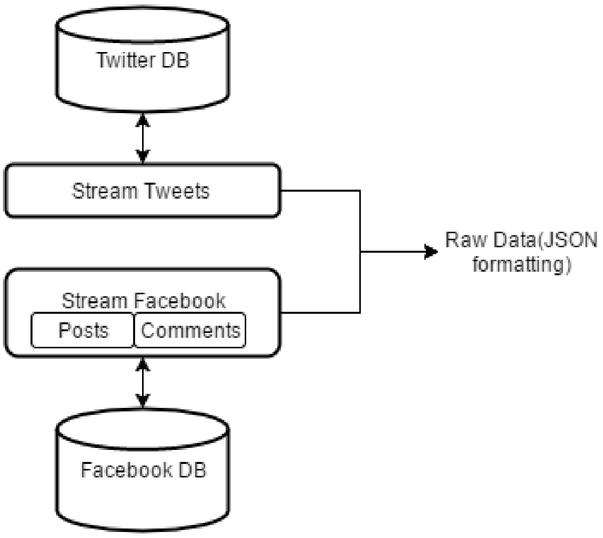


Figure 3.2 Data Collection for Text

Figure 3.3 shows the detailed design for the data collection submodule for numerics.

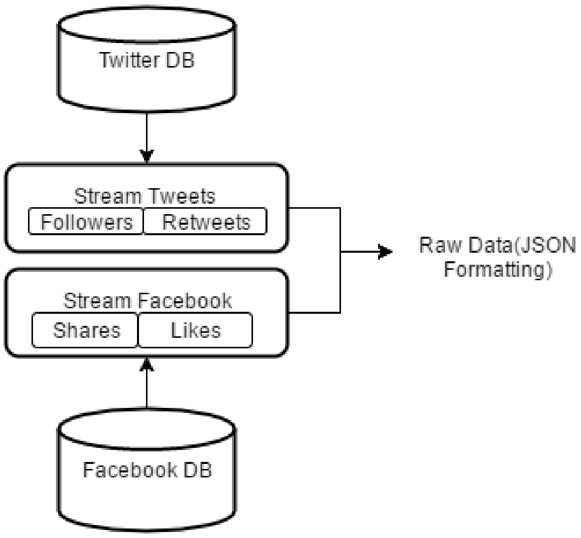


Figure 3.3 Data Collection for Numerics

Algorithm : Twitter/Facebook data extractor

```
get the necessary access tokens from Twitter developer site and
  insert into crawler
enter the desired brand name as keyword for crawling
extract tweets from the past 7 days
save the raw data which is of json format into a json file
for each array in the json object
        get the id tag
        get the name tag
        get the tweet tag
        get the retweet count tag
        get the follower count tag
save the extracted data into a csv file
```

3.4.2 Data Preprocessing

In this module, the posts and tweets crawled respectively from Facebook and Twitter are retrieved from the csv formatted file created in the previous module. Figure 3.4 shows the process flow for the data preprocessing module. The steps mentioned in the figure are as follows:

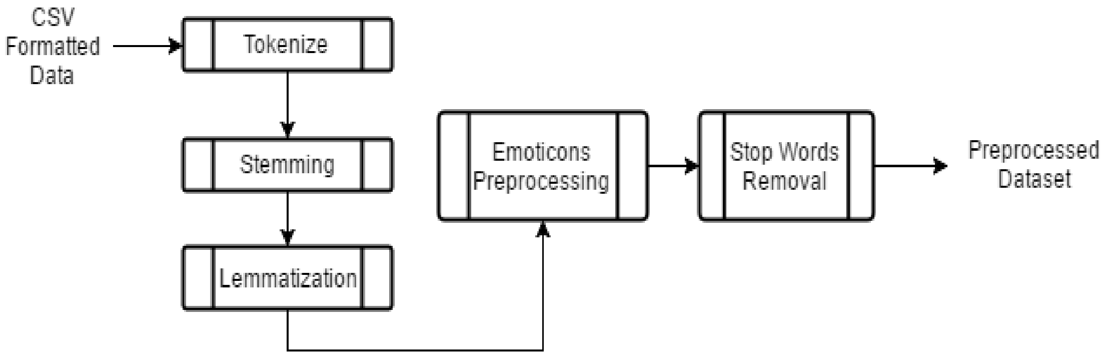


Figure 3.4 Preprocessing for Text Dataset

1. **Tokenization**

The text data retrieved from the csv is split into tokens. Paragraphs are split into sentences and sentences into words.

2. **Stemming**

It is the process in which the words are brought down to their root forms. Basically, the suffixes such as -ing, -er, -ed and so forth from the words. Porter Stemming algorithm is used for the above process. For example,

- playing to play
- started to start

3. **Lemmatisation**

This process is similar to stemming, instead it removes inflectional endings and return the base form of a word that is known as a lemma. For example,

- is, am, are to be
- customer's, customers, customers' to customer

4. **Emoticons preprocessing**

Emoticons are a pictorial representation of person's/user's feelings or emotions and it is written using special characters(!,:,),(,.), numbers and letters. The python script we have written identifies the emoticons and writes out the corresponding expression for the emoticon. For Example,

- :) to HAPPY

- :(to SAD

Algorithm : Emoticons Preprocessing

```
get the necessary access tokens from Twitter developer site
  and insert into crawler
enter the desired brand name as keyword for crawling
extract tweets from the past 7 days
save the raw data which is of json format into a json file
for each array in the json object
        get the id tag
        get the name tag
        get the tweet tag
        get the retweet count tag
        get the follower count tag
save the extracted data into a csv file
```

5. **Stop Words**

Words such as and, the, a, an do not add any meaning to the emotion of the sentences. Further, strings such as links will also be posted by the user in his/her tweets/posts which is not necessary for our analysis procedure. Hence these words and links which are of negligible use are removed using the stopword removal process.

3.4.3 Feature Extraction

Dataset

For text dataset, we manually crawled through Twitter and Facebook for tweets and posts that focused on user's experiences about an electronic product or brand. We filtered through the massive amount of data to retrieve only those that we felt were more expressive about the brand and its product. Figure 3.5 shows the process flow of labelling our text dataset. This was done as follows: AFINN dictionary - This is an English dictionary which consists of 2477 words which are given certain scores. These scores range from -5 to 5. We compared our text data with this dictionary and gave scores to each tweet/post based on the final scores from the code and hence labelled them according to the algorithm.

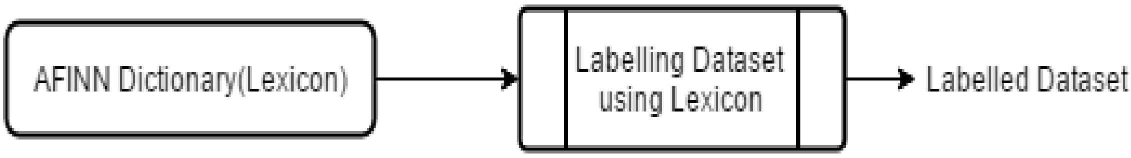


Figure 3.5 Labelling for Text Dataset

Algorithm : Labelling using AFINN dictionary

```
load the text data into a string
strip the sentences from the string
for sentence in string
    tokenize the sentences and store words in a list A
    load the AFINN dictionary into a list B
    for word in list A
       compare it with dlist B
       if match found
          increment the score by corresponding AFINN score
       else break
    get overall sentence score
       label the sentence based on the overall sentence score
write the sentences to a csv file according to the label
```

Hence the text dataset is created in the above format, that is, the sentences are placed in rows against their corresponding labels. Now, the dataset is ready to be trained by the Multiclass NaiveBayes classifier where each class is nothing but the labels given to the sentences.

For numerical dataset, again we crawled through Facebook and Twitter to retrieve key features for a particular brand/company such as

1. number of likes
2. number of shares
3. number of comments
4. retweet count
5. follower count

We had to retrieve an ample amount of data for our classification module. So the same procedure was done for another 100-200 companies. Our next step was to find the mean ,standard deviation and covariance of the retrieved 5 features and append to the csv file. Another 150 instances were appended to the final dataset using the same procedure. The covariance value is used to label the dataset of the four classes: A,B,C,D with A being the the highest grade in user engagement and user reach and D being the lowest grade. The equation for covariance can be referenced in 3.1.

$$\$cov_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1} \tag{3.1}$$

The likes and shares of a post contributed to the level of user reach as the more number of likes and shares means the post has reached more number of people ,whereas, user engagement corresponds to the user's perspective about the post and how involved he is on the topic. We then found the average of covariance of likes and shares and compared this with covariance of number of comments. Priority given to user engagement i.e; covariance of number of comments and the corresponding label was given for the brand based on a range of values. The labeled dataset is then passed to the SVM classifier.

3.4.4 Training and Testing

This module can be split again for text and numerical dataset. We are using the classification models in the ”Scikit-Learn” namely

- Multinomial Naive Bayes
- Support Vector Machine

Figure 3.6 shows the detailed design of the training and testing module.

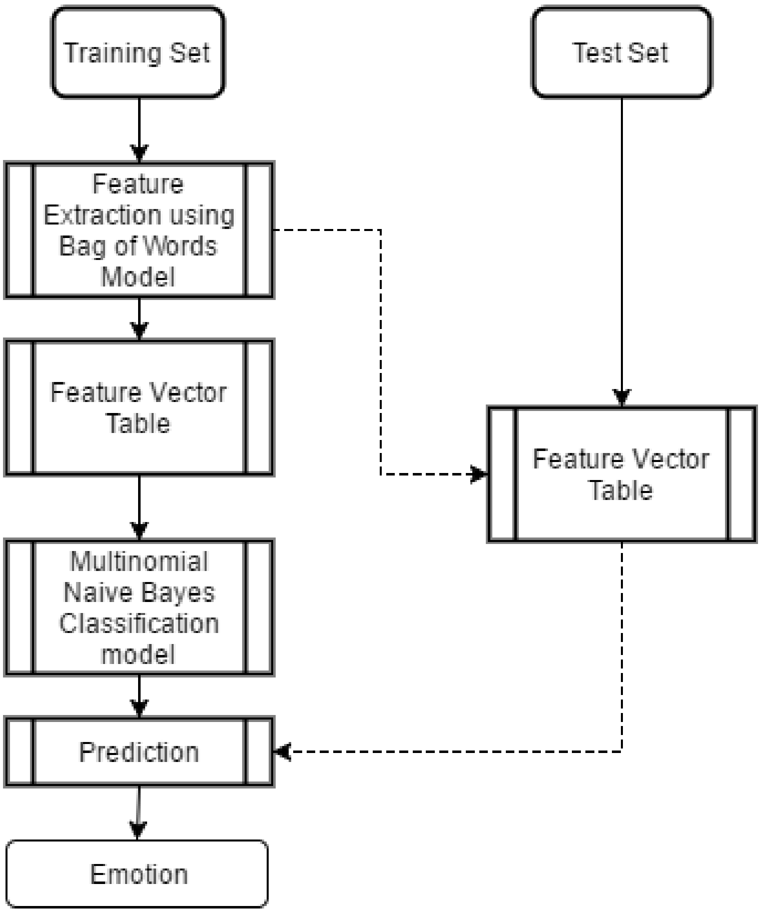

Training Set
Test Set
Feature Extraction using Bag of Words Model
Feature Vector Table
Feature Vector Table
Multinomial Naive Bayes Classification model
Prediction
Emotion

Figure 3.6 Training and Testing for Textual features

The training dataset is fed to the classifier retrieved from the previous module. For text dataset, the features are extracted using the Bag of Words model. The bag of words model is used in natural language processing and text mining applications. This model processes a text such as a sentence or a document is represented as a multiset or bag of words without considering grammar and word order but keeping multiplicity. The feature vectors are contructed using TF-IDF(Term Frequency - Inverse Document Frequency). Basically, the text frequencies noted above are down-weighted by the frequency of the words in corpus. This process is called vectorization. Figure 3.7 depicts the set of features after Feature Extraction. The following three tasks are carried out in the bag of words model:

1. Tokenizing
2. Counting occurence of tokens
3. Normalizing and weighting the tokens

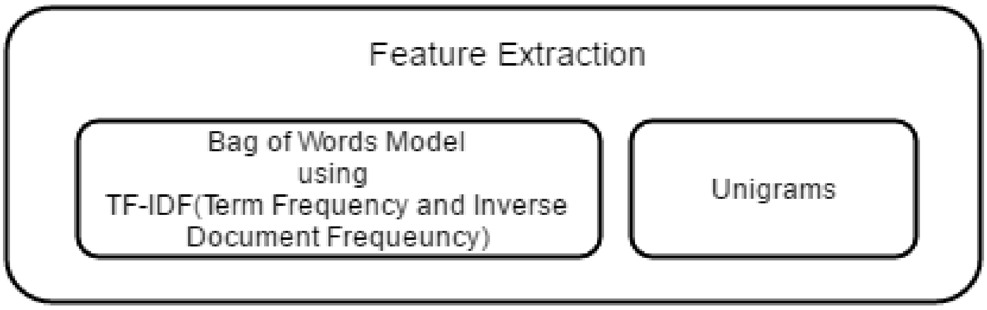

Feature Extraction
Bag of Words Model using TF-IDF(Term Frequency and Inverse Document Frequeuncy)
Unigrams

Figure 3.7 Features for Text

Ngram features are found to be useful for text classification tasks. We have used unigrams as features for emotion identification. Unigrams are found to be very useful features and these include adverbs, verbs, adjectives and nouns. The Scikit-Learn package includes predefined functions to extract the ngram features in a corpus.

The features are then fed to the classifier for training the model. The dataset is split into training set and test set where 3/4th is weighted for training and 1/4th is given as test.

Naive Bayes Classifier for Multinomial Models

The MNB classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. It works well against multiclass classification. We have seven emotion labels namely:

- Anger
- Joy
- Sadness
- Guilt
- Fear
- Shame
- Disgust

Denoting the classes as C_i and any relevant features F_i, the probability of a given class is given by Bayes Theorem referenced in 3.2,

$$p\left(C_i|\vec{F}\right) = \frac{p\left(\vec{F}|C_i\right)p(C_i)}{\sum_j p\left(\vec{F}|C_j\right)p(C_j)}. \tag{3.2}$$

Now all we need to do is model the class likelihoods, $p(\vec{F}|C_i)$. The Naive Bayes assumption is that the features are independent given a class, i.e. $p(\vec{F}|C_i) = \prod_j p(F_j|C_i)$. Popular choices of the $p(F_j|C_i)$. include the Bernoulli distribution (taking into account whether a binary feature occurs or not) and the Binomial distribution (taking into account not just the presence of a binary feature but also its multiplicity).

When a feature is discrete but not binary (or continuous but approximated by such discrete values) a common choice of likelihood is the multinomial distribution, hence Multinomial Naive Bayes.

Multiclass SVM

Support vector machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression. In this algorithm, we plot each data item in our set in an n-dimensional space where n is the number of features with the value of each feature being the value of a particular coordinate. Classification is then done by finding the hyperplanes that separates two classes (in our case 4 classes). Figure 3.8 shows the process flow of the training and testing submodule for numerical features.

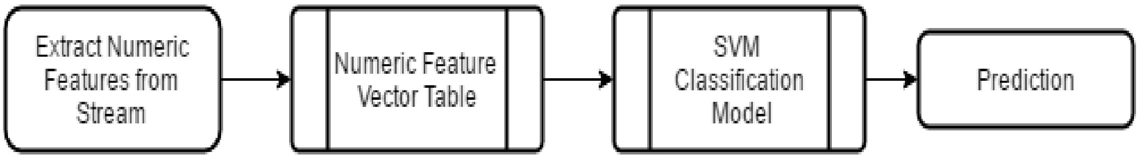

Extract Numeric Features from Stream
Numeric Feature Vector Table
SVM Classification Model
Prediction

Figure 3.8 Training and Testing for Numerical features

3.4.5 Real Time Crawling and Data Visualisation

Once the training is done, we have to test our system with the real time data which are being posted by the users currently. This requires the crawling and testing of data simultaneously. This has been done as follows:

Algorithm : Text Data Testing

```
insert the access keys and tokens required to stream Twitter
enter the brand name which has to be crawled
while counter less than 100:
    save the tweets into a list T
    increment counter
for tweets in T:
    remove stop words from the tweets
    remove links and unwanted spaces using regular expressions
    store cleaned data in list new_T
        bow_scores = word scores using bag of words transform of new_T
        class_name = predicted class using bow_scores
    print class_name, tweets
store output in a text file
```

Algorithm : Numerical Data Testing

```
insert the page access token and api keys
enter the name of the page as on Facebook
while counter less than 100
    collect the likes, comments and shares of each post
    store data in corresponding columns in csv file A
open A
for each column in A
    calculate mean of each column
    calculate standard deviation for each column
    store the values in list A
class = predicted class using elements in list A
print class
```

Data Visualisation

Once the real time crawling and testing is completed, we need to visualise these results for easier understanding. We have included the following plots based on our test results:

Heat map based on location : We were able to get the location of the users along with their tweets . We used these location details to find the density of the number of users tweeting from a particular region and generated a heat map based on our results.

Pie chart based on the classified emotions : Since we have 7 emotion classes in our classification process, we picturised the classification based on these emotions on a piechart. Each colour on the piechart represents an emotion based on the number of tweets falling under that particular emotion.

Bar graph : We have plotted a bar graph based on the number of likes, comments and shares from facebook and the number of tweets that fall under each emotion.

CHAPTER 4

IMPLEMENTATION AND RESULTS

This chapter deals with experimental analysis of the proposed system.

4.1 DATASET DESCRIPTION

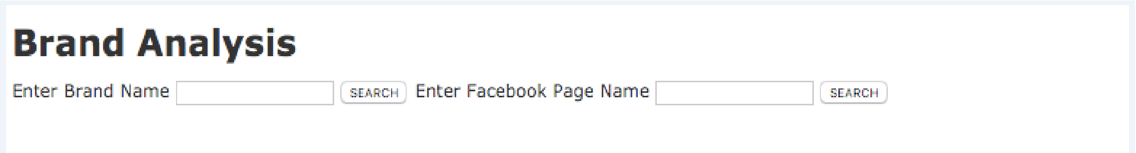
The user must give a single input to the crawler.

- Keyword(Brand name)

The crawler runs through Twitter and Facebook to search the keyword and spews out JSON formatted data. It is then converted to CSV format with the required tags. Various keywords like Microsoft, Huawei, Samsung were given as input.

4.2 EXPERIMENTAL RESULTS

Figure 4.1 shows the user interface. The user interface gets the keyword/brand name from the user.

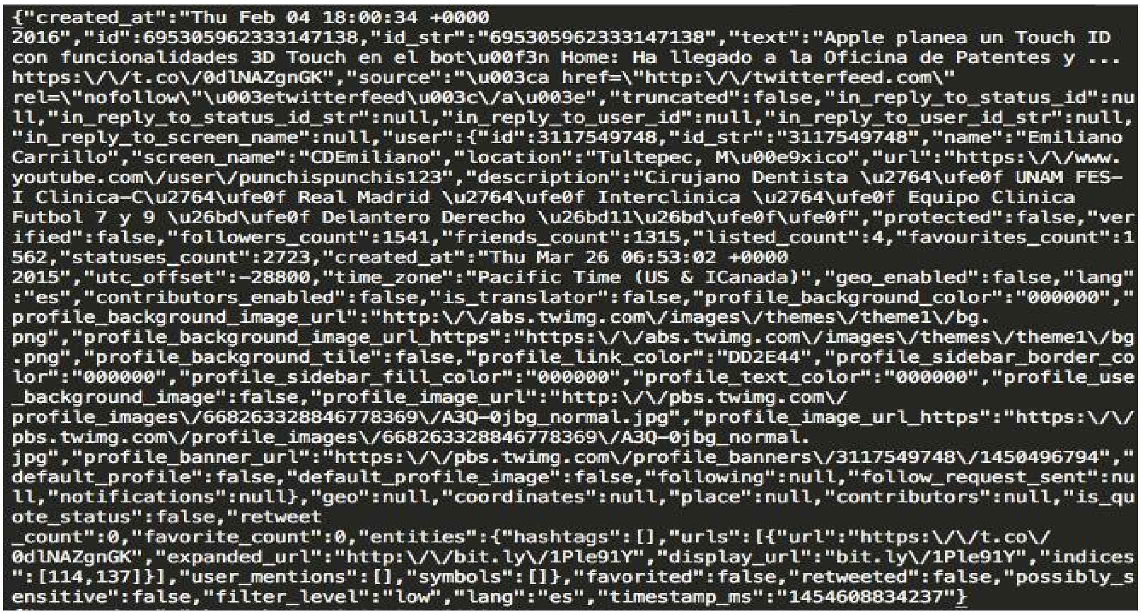

Brand Analysis

Enter Brand Name

Enter Facebook Page Name

Figure 4.1 User Interface

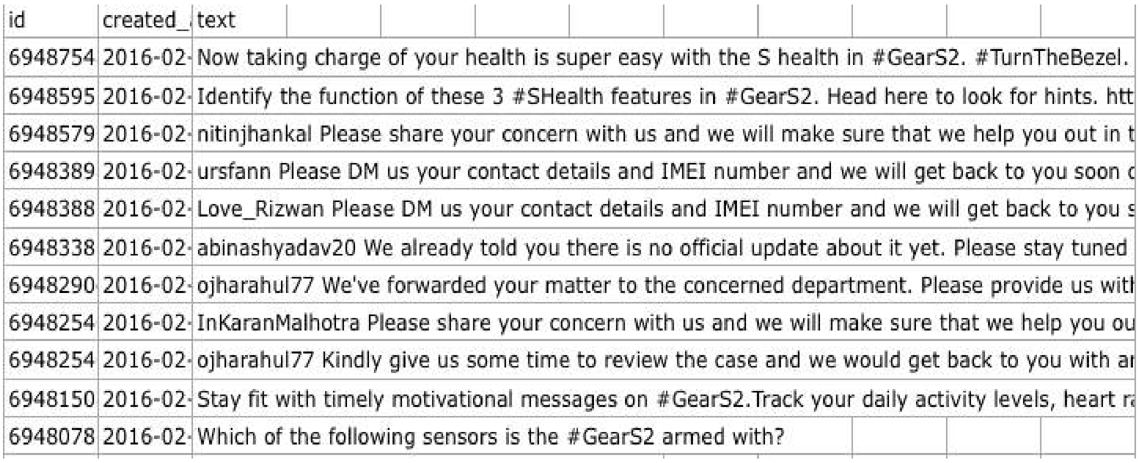
Figure 4.2 shows the retrieved data from Facebook. It is structured in JSON format.



```
{"created_at":"Thu Feb 04 18:00:34 +0000
2016","id":695305962333147138,"id_str":"695305962333147138","text":"Apple planea un Touch ID
con funcionalidades 3D Touch en el bot\u00f3n Home: Ha llegado a la Oficina de Patentes y ...
https:\/\/t.co\/0dlNAZgnGK","source":"\u003ca href=\"http:\/\/twitterfeed.com\"
rel=\"nofollow\"\u003etwitterfeed\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":nu
ll,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,
"in_reply_to_screen_name":null,"user":{"id":3117549748,"id_str":"3117549748","name":"Emiliano
Carrillo","screen_name":"CDEmiliano","location":"Tultepec, M\u00e9xico","url":"https:\/\/www.
youtube.com\/user\/punchispunchis123","description":"Cirujano Dentista \u2764\ufe0f UNAM FES-
I Clinica-C\u2764\ufe0f Real Madrid \u2764\ufe0f Interclinica \u2764\ufe0f Equipo Clinica
Futbol 7 y 9 \u26bd\ufe0f Delantero Derecho \u26bd11\u26bd\ufe0f\ufe0f","protected":false,"ver
ified":false,"followers_count":1541,"friends_count":1315,"listed_count":4,"favourites_count":1
562,"statuses_count":2723,"created_at":"Thu Mar 26 06:53:02 +0000
2015","utc_offset":-28800,"time_zone":"Pacific Time (US & ICanada)","geo_enabled":false,"lang"
:"es","contributors_enabled":false,"is_translator":false,"profile_background_color":"000000","
profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.
png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg
.png","profile_background_tile":false,"profile_link_color":"DD2E44","profile_sidebar_border_co
lor":"000000","profile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use
_background_image":false,"profile_image_url":"http:\/\/pbs.twimg.com\/
profile_images\/668263328846778369\/A3Q-0jbg_normal.jpg","profile_image_url_https":"https:\/\/
pbs.twimg.com\/profile_images\/668263328846778369\/A3Q-0jbg_normal.
jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/3117549748\/1450496794","
default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":nu
ll,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_qu
ote_status":false,"retweet
_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/t.co\/
0dlNAZgnGK","expanded_url":"http:\/\/bit.ly\/1Ple91Y","display_url":"bit.ly\/1Ple91Y","indices
":[114,137]}],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"possibly_s
ensitive":false,"filter_level":"low","lang":"es","timestamp_ms":"1454608834237"}
```

Figure 4.2 Retrieved Data in JSON format

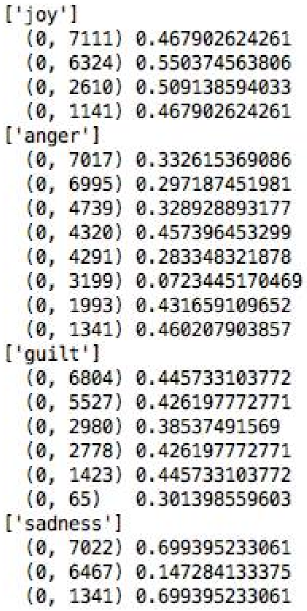
Figure 4.3 shows the same data in CSV format after preprocessing. The preprocessed data is fed to the feature extractor to retrieve important features.



id	created_	text
6948754	2016-02-	Now taking charge of your health is super easy with the S health in #GearS2. #TurnTheBezel.
6948595	2016-02-	Identify the function of these 3 #SHealth features in #GearS2. Head here to look for hints. htt
6948579	2016-02-	nitinjhankal Please share your concern with us and we will make sure that we help you out in t
6948389	2016-02-	ursfann Please DM us your contact details and IMEI number and we will get back to you soon c
6948388	2016-02-	Love_Rizwan Please DM us your contact details and IMEI number and we will get back to you s
6948338	2016-02-	abinashyadav20 We already told you there is no official update about it yet. Please stay tuned
6948290	2016-02-	ojharahul77 We've forwarded your matter to the concerned department. Please provide us witl
6948254	2016-02-	InKaranMalhotra Please share your concern with us and we will make sure that we help you ou
6948254	2016-02-	ojharahul77 Kindly give us some time to review the case and we would get back to you with ar
6948150	2016-02-	Stay fit with timely motivational messages on #GearS2.Track your daily activity levels, heart ra
6948078	2016-02-	Which of the following sensors is the #GearS2 armed with?

Figure 4.3 Preprocessed Data in CSV Format

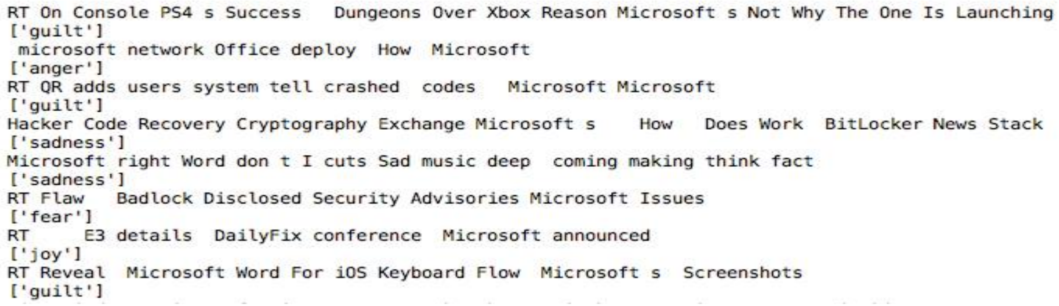
Figure 4.4 shows the list of features and their term weightage values. This step is for feature extraction.



```
['joy']
  (0, 7111) 0.467902624261
  (0, 6324) 0.550374563806
  (0, 2610) 0.509138594033
  (0, 1141) 0.467902624261
['anger']
  (0, 7017) 0.332615369086
  (0, 6995) 0.297187451981
  (0, 4739) 0.328928893177
  (0, 4320) 0.457396453299
  (0, 4291) 0.283348321878
  (0, 3199) 0.0723445170469
  (0, 1993) 0.431659109652
  (0, 1341) 0.460207903857
['guilt']
  (0, 6804) 0.445733103772
  (0, 5527) 0.426197772771
  (0, 2980) 0.38537491569
  (0, 2778) 0.426197772771
  (0, 1423) 0.445733103772
  (0, 65)   0.301398559603
['sadness']
  (0, 7022) 0.699395233061
  (0, 6467) 0.147284133375
  (0, 1341) 0.699395233061
```

Figure 4.4 Feature Extraction

Figure 4.5 shows the tweet/post and its emotion after classfication. The output is displayed on the python console.



```
RT On Console PS4 s Success   Dungeons Over Xbox Reason Microsoft s Not Why The One Is Launching
['guilt']
 microsoft network Office deploy  How  Microsoft
['anger']
RT QR adds users system tell crashed  codes   Microsoft Microsoft
['guilt']
Hacker Code Recovery Cryptography Exchange Microsoft s    How   Does Work  BitLocker News Stack
['sadness']
Microsoft right Word don t I cuts Sad music deep  coming making think fact
['sadness']
RT Flaw   Badlock Disclosed Security Advisories Microsoft Issues
['fear']
RT     E3 details  DailyFix conference  Microsoft announced
['joy']
RT Reveal  Microsoft Word For iOS Keyboard Flow  Microsoft s  Screenshots
['guilt']
```

Figure 4.5 Prediction output on Console

Figure 4.6 shows the data visualisation module. The output is displayed on the web browser. The module displays pie chart depicting the distribution of emotions from the real-time dataset, bar graph depicting the number of likes, shares and comments i.e, user reach and user engagement and finally a heat map based on location.

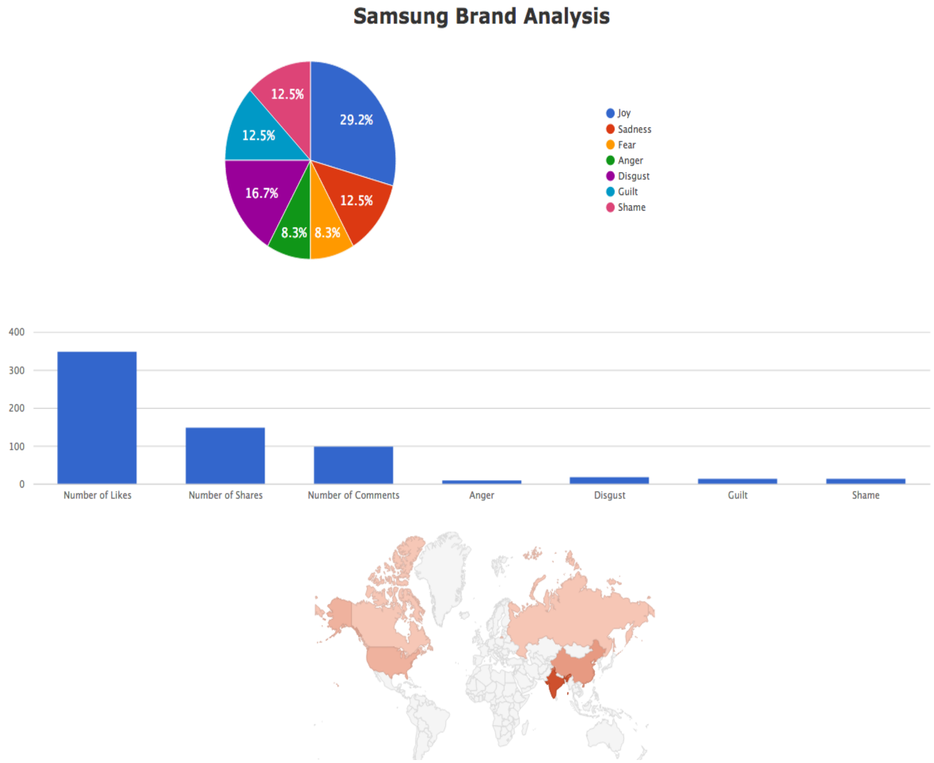


Figure 4.6 Data Visualisation

4.3 PERFORMANCE EVALUATION

We use three parameters to evaluate the project:

1. Recall - It is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate.

$$Recall = \frac{TP}{TP+FN}$$

2. Precision - It is the number of positive predictions divided by the total number of positive class values predicted. It is also called the Positive Predicted Value(PPV).

$$Precision = \frac{TP}{TP+FP}$$

3. F-Score - A measure that combines precision and recall is the harmonic mean of precision and recall.

$$F = 2.\frac{precision.recall}{precision+recall}$$

Table 4.1 shows the precision and recall for the MNB classifier. The corresponding comparison is shown via a bar chart in Figure 4.7 and 4.8

Emotion Label	Precision	Recall	F-Score
Anger	0.72	0.67	0.72
Joy	0.74	0.82	0.78
Disgust	0.84	0.67	0.75
Sadness	0.73	0.74	0.74
Shame	0.70	0.70	0.70
Guilt	0.66	0.73	0.70
Fear	0.77	0.76	0.77

Table 4.1 Precision and Recall for Multinomial Naive Bayes Classifier

Table 4.2 shows the precision and recall for the SVM classifier. The corresponding comparison is shown via a bar chart in Figure 4.9 and 4.10

Class	Precision	Recall	F-Score
A	74	76	72
B	71	73	68
C	65	68	67
D	59	62	60

Table 4.2 Precision and Recall for SVM Classifier

Figure 4.7 is a relationship between the emotion and the precision values for each emotion. Precision (for each emotion) in this case tells us the fraction of retrieved words from the entire dataset, which are relevant and fall under that particular emotion. As we can see from the graph, ”disgust” has the highest precision value and ”guilt” has the lowest precision value indicating that, the number of relevant words falling under the ”disgust” class is the highest where as the ”guilt” gets the least number of relevant words.

As far as precision is concerned, only the relevance of each class can be found. In order to find the fraction of the correct words in the relevant set, recall values for each emotion were calculated. The relationship between the emotion and its' recall values have been plotted in Figure 4.8. ”joy” has the highest recall value where as anger has the lowest.

The precision and recall values for the Support Vector Machine classifier are also found similarly and plotted in the following graphs. This is referenced in 4.9 and 4.10

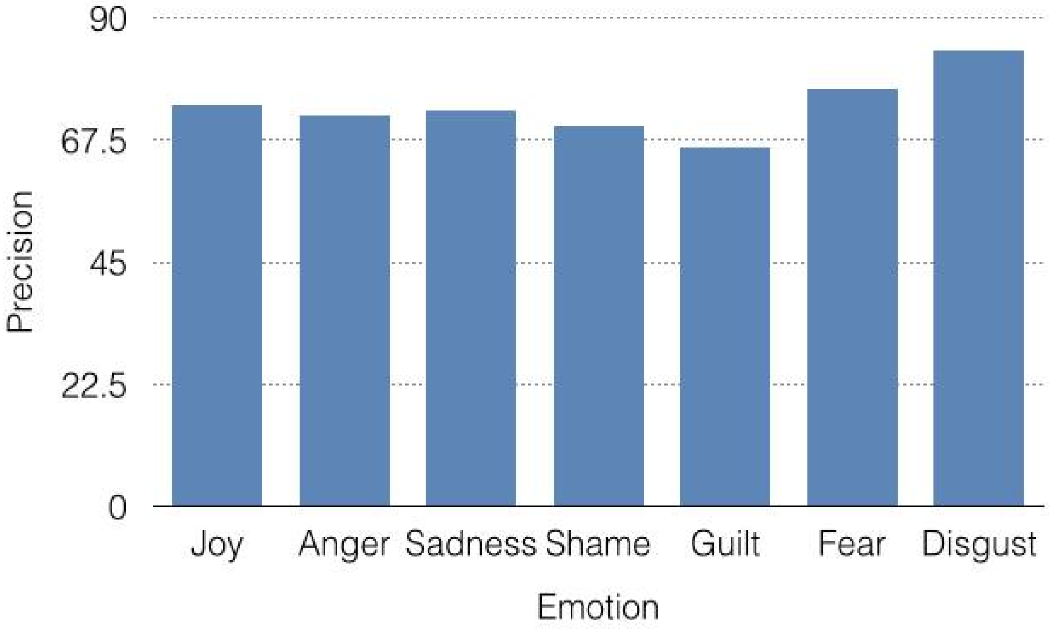


Figure 4.7 Precision for Multinomial Naive Bayes Classifier

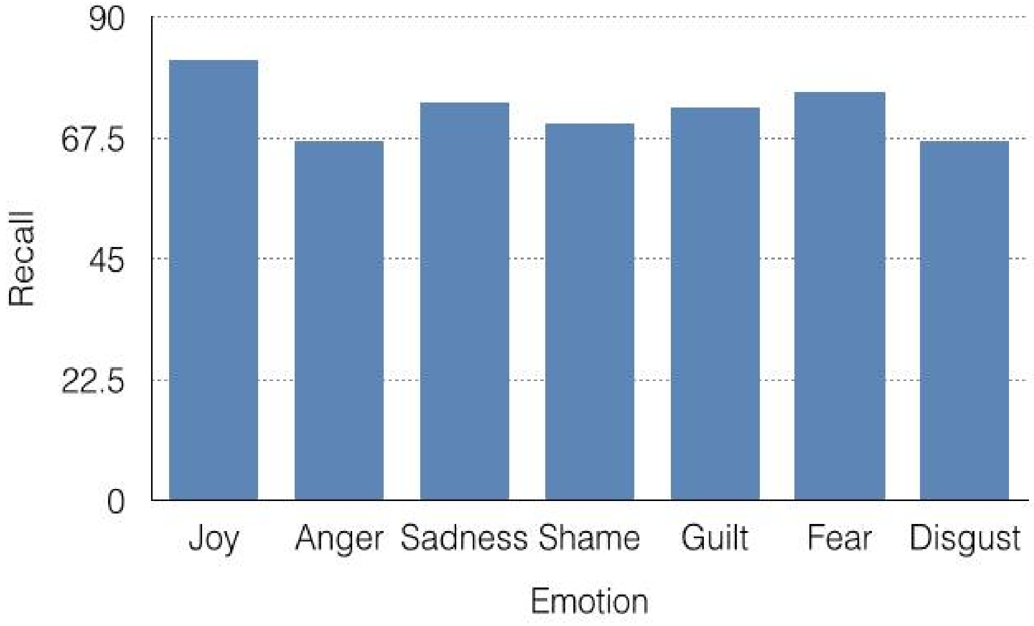


Figure 4.8 Recall for Multinomial Naive Bayes Classifier

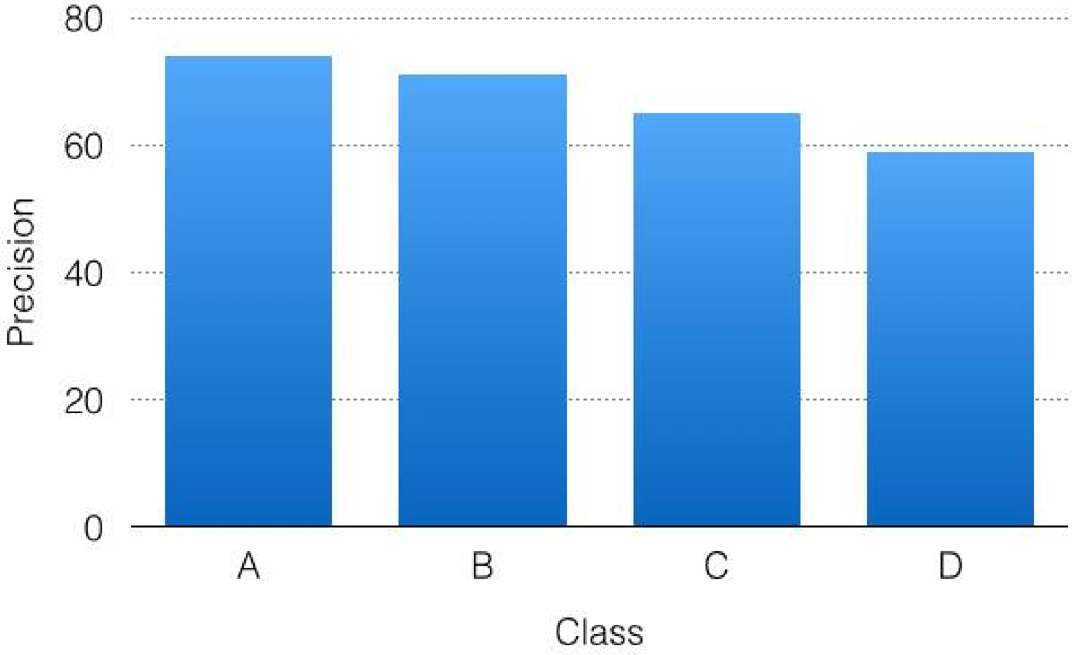


Figure 4.9 Precision for SVM Classifier

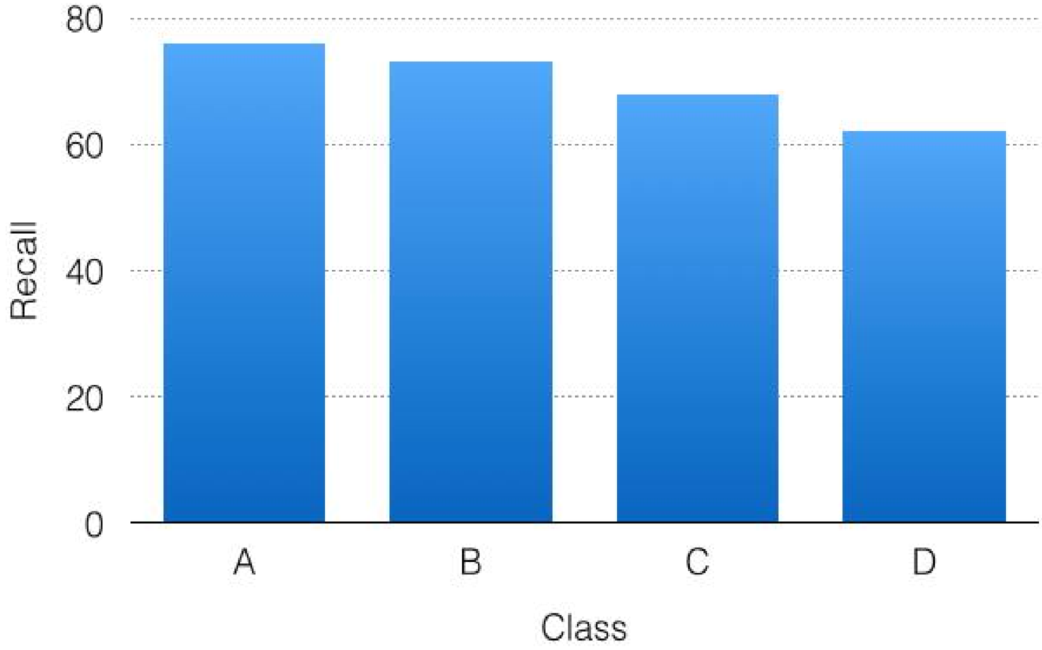


Figure 4.10 Recall for SVM Classifier

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

A brand evaluation system has been developed that includes an user interface to accept the keyword(the brand name) and crawls through popular social media sites Facebook and Twitter. Only those posts and tweets which refer to the brand name are returned. The returned results are very accurate. The results are filtered and converted to a readable format and it is then fed to the preprocessing module. The preprocessing module tokenizes, removes inflections and stop words such as URLs, articles and so forth from the dataset. The preprocessed dataset is sent to the classifiers module that predicts the emotion of tweets, posts and comments from Twitter and Facebook respectively and also predicts the level of user reach and user engagement.

A web app has been developed that displays pie charts, graphs, and heat maps based on the results. Further analysis on the graphical data will reveal more important features that improve the marketing aspects of the brand on social media sites.

5.2 FUTURE WORK

Although Social Media monitoring had evolved drastically in the past couple of years, the future has a lot in store. The data collection and analytics tools will grow both in features and in number. Accuracy will skyrocket due to advancement in artificial intelligence and better algorithms. There will also be more data to be collected as every day more and more people join the social networks and start posting their opinions on brands, events, and their life.

With Facebook's recent update of changing the ever so popular like button to the new reaction button which features the emotions of like, love, haha, yay, wow, sad, angry, we will be able to delve much deeper into how a customer or a potential customer feels about a brand and its products. This insight will help brands and companies grow their strengths and shield their weaknesses.

Damage control will become a lot easier, yet a lot more time-consuming as instead of issuing a generic message to the media of the masses like the newspaper, the television amongst others, reaching out to each and every individual and personally apologising will become the norm if a brand wishes to keep their position in the market.

In the near future, we could also expect monitoring images instead of just text. With image recognition developing fast, we can expect social media monitoring tools to keep and eye on the pictures that people are uploading as well. This will help the brands know who is using their product and thus help them discover who to target for up-sell and likes.

Video is also growing at a rapid rate. Although it may take sometime, it would not be surprising if social media monitoring tools starting taking advantage of the large amounts of video available on the web. This would also help monitor the ever so popular unboxing and review videos that are all over youtube and other video sites.

REFERENCES

[1] “Afinn dictionary”, http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010. Accessed: 2011.

[2] “Creen project”, http://www.creen.org. Accessed: 2011.

[3] Mark E Larsen, Tjeerd W Boonstra, Philip J Batterham, Bridianne ODea, Cecile Paris, and Helen Christensen, “We feel: mapping emotion on twitter”, *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, num. 4, pp. 1246–1252, 2015.

[4] Dongjoo Lee, Ok-Ran Jeong, and Sang-goo Lee, “Opinion mining of customer feedback data on the web”, In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pp. 230–235. ACM, 2008.

[5] Alexander Pak and Patrick Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining.”, In *LREc*, volume 10, pp. 1320–1326, 2010.

[6] Carlo Strapparava, Alessandro Valitutti, et al., “Wordnet affect: an affective extension of wordnet.”, In *LREC*, volume 4, pp. 1083–1086, 2004.

[7] Bincy Thomas, P Vinod, and KA Dhanya, “Multiclass emotion extraction from sentences”, *International Journal for Scientific and Engineering Research (IJSER)*, vol. 5, num. 2, pp. 12–15.