

MAPPING SENTIMENTS VIA TWEETS



FINAL YEAR PROJECT UG 2013

By

Leader – 08268 Muhammad Alqamah Bhutta

Member 1 – Suliman Qais Khan

Member 2 – Mir Balach Marri

Member 3- Saifullah Khan

NUST Institute of Geoinformatics Engineering
School of Civil and Environmental Engineering
National University of Sciences and Technology, Islamabad,
Pakistan

2017

This is to certify that the

Final Year Project Titled

MAPPING SENTIMENTS USING TWEETS

submitted by

Leader – 08268 Muhammad Alqamah Bhutta

Member 1 – Suliman Qais Khan

Member 2 – Mir Balach Marri

Member 3- Saifullah Khan

Has been accepted towards the requirements

For the undergraduate degree

in

GEOINFORMATICS ENGINEERING

Ms Quratulain Shafi

Lecturer

NUST Institute of Geoinformatics Engineering

School of Civil and Environmental Engineering

National University of Sciences and Technology, Islamabad, Pakistan

TWEET MAPS

ABSTRACT

In the era of big data and virtual reality lack of Geographic Information Systems to integrate new type of data and visualization methods is limiting its applications and full potential, in order to enhance GIS capabilities. This project provides new methods of gathering data from crowd sourcing and social media and helps display it in a compact and visually appealing method to improve decision making and analysis. This system extracts information from tweets by data mining and computer intelligence, then displays them categorically using maps. Furthermore users can get sentiment analysis of society for information of their interest. By utilizing information from social media a new dimension has opened up for analysis using GIS.

DEDICATED TO

My mother who continues to see the good and has encouraged and helped us complete this task.

ACKNOWLEDGEMENT

We wish to thank our mentor, Ms Quratulain Shafi for her idea of extracting sentiment analysis from tweets.

TABLE OF CONTENTS

ABSTRACT

DEDICATION

ACKNOWLEDGEMENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODCUTION1

 Background

 Objective

 Study Area

 Scope

 Twitter

CHAPTER 2

METHODOLOGY1

 Repository

 R Stuido

 Tweepy

 Hashtags and Followers

 Tweets

 Data cleaning

 Information Extraction and Transformation

 Geocoding and Geofencing

 Classification

 Classes

 Naïve Bayers Classifier

 Sample data

 Training Classifier

 Optimization

 Sentiment Analysis

 Web Architecture

 Integration

 Visualization

 Real-time mapping

CHAPTER 3
RESULTS AND DISCUSSION

Working
Connections and issues
Trends
Visual Analysis
Sentiments
Discussions

CHAPTER 4
CONCLUSION AND RECCOMENDATION

APPENDICES

BIBLIOGRAPHY

LIST OF FIGURES

Figure 1: Probability method	18
Figure 2: Learning curves	21
Figure 3: Web front	27
Figure 4: General Tab	28
Figure 5: Sentiment Tab	29

LIST OF TABLES

Table 1: Classifier

21

CHAPTER 1

INTRODUCTION

Social media consists of online interactive websites and mobile applications populated by user generated content in the form of texts, posts, pictures, video, tweets and comments.

Some of the most common social media sites are Facebook, Twitter, Google+ and Reddit. The data on these websites contains information regarding user feelings, trends, hypes and ongoing issues.

The major source of GIS data consists of remotely sensed satellite or aerial images and previously existing maps. Although very detailed analysis of earth can be performed using this data, yet in the current era of big data where spatial data does not consist only in the form of raster (pixels) or vector analytical capabilities can be significantly enhanced if information is extracted from social media applications.

Cartography and visualization is a major aspect of GIS and is the major way users can quickly assess geographic information and study trends.

This project aims to enhance the domain of GIS by incorporating new data sources, using modern methods to extract meaningful information from them and displaying them using new technologies such as virtual reality and interactive maps

1.1 Background

Greeks made the first map in 5 BC using their own symbols. The field of cartography developed gradually from there forming basic elements of map and rules for displaying information. GIS were developed in 20th centenary where they were used to solve adhoc problems relating to earth and various spatial phenomenons. In the last 4 decades GIS

has developed rapidly as satellite imagery became more popular and computing powers increased.

In the new age simple maps are becoming out dated and market users require new ways to gather information. GIS is continuously developing and has an inherent capability of transforming data. By utilizing new technology and incorporation information from sea of data available at social media websites information will be displayed much faster and effectively.

1.2 Objective

The objective of this project is to “create an interactive web platform to display geospatial information gathered from twitter and provide sentiment analysis using data mining and geographic information system principles.”

The objective was subdivide into four main categories

- i. Download tweets and extract required information
- ii. Build neural network to keep website up dated on latest trends
- iii. Perform sentiment analysis on tweets
- iv. Make web interface to show information in various formats

1.3 Study Area

The scope of this project is limited to Pakistan and topics affiliated to Pakistan, such as news channels, celebrities, politicians and trends originating in Pakistan. Although any tweet which has information associated with Pakistan even though its origin is outside the country will be shown with its true location

1.4 Scope

Since there are numerous social media websites out there with billions of users each having its own format and associated information, copyrights and regulations it was necessary to define the scope of this project to limit various formats of data and to maintain user privacy agreements.

Keeping this in mind we chose twitter to be our main source of data and limited its region to Pakistan. For visualization we set online maps to be our base and categorized tweets to increase user friendliness.

1.5 Twitter

Twitter is an online website where users can post information in the form of tweets. A tweet is at max 140 characters long and contains any information that the user intends to convey. Tweets are associated to users and hash tags, by following a particular user you get to see all their tweets while hash tags are a general categorization of tweets which helps define the topic of tweet. Tweets can be shared by users a process known as retweeting, the number of times a tweet has been shared shows its popularity.

We used tweets because they have the following benefits

- 1: Tweets are short and very precise; most tweets have a particular agenda or topic. Information is in the form of texts so text classification can be utilized to extract meaningful information
- 2: Twitter has a user agreement which allows users to gather tweets from different people and utilize them as they want without fear of infringement of privacy.
- 3: Most tweets are geo-tagged thus they are associated to a particular location on earth. Thus their location can easily be extracted unlike information from other social platforms

CHAPTER 2

METHODOLOGY

2.1 Repository

The first step to making this project was to form a central repository which would serve as the base for gathering information from twitter. The repository will consist of hash tags and follower ids which will be used to download tweets. In order to build a repository we had to find out about the most influential people and famous twitter topics related to Pakistan.

2.1.1 R Studio

R studio is an integrated development environment for R. It uses a large user community which helps develop libraries and codes. R is an open source programming language and software environment for statistical computing and graphical display. It is widely used for data mining and statistical analysis by data analysts.

We used R to continuously download tweets from twitter, from these only those were kept which contained information linked to Pakistan either directly or indirectly. Once we had these tweets we needed a criteria to evaluate which topics the users of Pakistan were interested in, to do this number of retweets was set to be the determining criteria. 100 of the most famous hash tags and user accounts were selected from this dataset to serve as our base repository.

2.1.2 Tweepy

Once a suitable repository was established we needed a platform to download tweets, for this purpose python was used with an online library known as “Tweepy”. Tweepy allows users to download tweets in real-time by providing geographical region, hash tags or user accounts.

This data is in JSON format and contains information contains tweet, hash tags, user information, user address and location of tweet. Sometimes information is missing so it needs to be extracted from hash tag or the tweet itself.

2.1.3 Hashtags and Followers

Hashtags are user defined topic, they are made arbitrarily by the person making the tweet. They start from the preliminary character # for example **#Pakistan**. Sometimes hash tags become famous and users start associating them with particular topics and using them in their tweets to show what they are talking about. A given tweet may contain more than one hash tag. Although these hashtags can serve as a basis for determining the topic of the tweet, they can still be deceptively wrong as users may make them up or use them out of context. In our project hashtags served as a base source for gathering tweets in the preliminary stage.

People who make twitter accounts form its user community. Users can follow other users to see the tweets they post, mostly famous personalities and organizations are followed by a large community base i.e. followers. Following a particular user you can get information that they post and this information can be of considerable importance to society if the user is someone important like President Donald Trump or IMF, but the downside is that a user may post tweets about various topics.

2.2 Tweets

Tweets are user thoughts and feelings about anything in the form of text. Recently tweets have become very famous and almost all public icons use tweets to give public information they want directly.

2.2.1 Data Cleaning

A downloaded tweet contains information in JSON format. The information contained is not just the users text but also various other associated data. One issue with data gathered from online sources is that it contains a lot of excess characters, miscellaneous texts and other errors. Thus tweets needs to be cleaned in order to make the information reliable and usable.

To clean tweets we studied the patterns of downloaded tweets and found the type of anomalies associated with each tweet. These anomalies were then removed by developing text classification algorithms using segmentation, loops and natural language text classification.

2.2.2 Information Extraction and Transformation

Once tweets were cleaned the next major step was to extract various types of information in tweets which were required for our platform. This included:

- Tweet: The text that user intended to convey
- Hashtags: The associated hashtags that the user included in his tweet
- User account: The name of user who posted this tweet
- Retweets: The number of times a tweet has been shared by users
- Location: The place where the user was when he posted his tweet
- Time: The time when the tweet was posted.

2.2.3 Geocoding and Geofencing

Tweets can utilize cell phone GPS or extract location of laptop using LAN address, but sometimes users have their location services turned off. In order to display tweets on map location was a key component. To overcome this difficulty we used the following methods in the given order or priority

1. Extract location if GPS was turned on.

If the user posted the tweet with his laptop or his cell phone when the location services are turned on, the downloaded tweet contains longitude and latitude of the users location in world geodetic system 1984 (WGS 1984). This information is extracted from the tweet and stored in the database along with the tweet.

2. Using text to determine location of tweet

If location services are turned off, the text of tweet is analyzed to find key words such as place names which are then used to extract the location of the tweet by using **GEOPY** library with the key words.

3. Hashtags to determine user location

If the tweet location cannot be extracted from the text then the linked hashtags are used as they may contain place names. This is not preferred as a person may be tweeting from one place but have hashtags of a completely different region. For example someone living in Karachi may be criticizing a riot in Islamabad and have the associated hashtag #Islamabad.

2.3 Classification

A major part of this project was to classify tweets into different categories to help users find their topics of interest.

2.3.1 Classes

We divided tweets into four main categories

1. Food

This category contains tweets associated to foods, restaurants and cafes etc.

2. Politics

It is a major topic in Pakistan media and is of high interest to people. It contains tweets about new political movements, political parties, tweets from political leaders and various other domains associated to the country's political situation.

3. Security

Tweets which contain information associated to terrorism, defense, security concerns and crimes are placed into this category.

4. Recreation

Visiting places, tourist spots, sports places and other sites of recreation are classified into it.

5. General

Since topics of tweets are very wide spread and sometimes there are trends which do not fit into the above categories, they are classified into this domain.

2.3.2 NLTK and RE

Regular Expression (RE) is a python module which is used for matching strings. RE was used to first remove any characters from strings such as backslashes, hashtags and miscellaneous words which contained special characters that might cause errors in classification. Then common articles of speech such as "a" , "of" etc were removed as the classifier might consider them important part of sentence and it would damage our

results. Finally non key words were removed from the tweet to optimize the classifier in order to save processing time and get quicker results.

Natural language toolkit (NLTK) is a library which is used for textual analysis. In our project we used NLTK to first optimize text in tweets by removing prefixes and suffixes, converting all texts into the same format so the classifier does not consider them different for classification and then used classifiers to separate text.

2.3.3 Naïve Bayes Classifier

This is a classifier based on Bayes theorem which uses probability to classify features while assuming strong independence among features.

The diagram shows the equation for Bayes' theorem: $P(c | x) = \frac{P(x | c)P(c)}{P(x)}$. Blue arrows point from the labels to the corresponding parts of the equation: 'Likelihood' points to $P(x | c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c | x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

FIGURE 1

In 2009 Alec Go and his colleagues performed sentiment analysis on tweets by using cluster of emoticons to represent positive and negative sentiments. They used a variety of classifiers and Naïve bays classifier gave the best results with an accuracy of 81% on their test results.

Since tweets varied greatly with respect to the information they contain we had to optimize the classifier by segmenting the tweet and applying max likely hood on different segment results and feature count in our data base. This improved the classification accuracy of our tweets.

2.3.3.1 Sample data

The first step to making a classifier is to gather sample data, in our case we downloaded tweets and manually separated them into categories we deemed suitable. These tweets were then segments and important features were extracted using the same function that the classifier will use.

An additional list of features was formulated by searching for key words associated with each category and refining the list to remove anomalies. Finally similes of each word were also included to further improve our feature list.

2.3.3.2 Training Classifier

Once the sample data was formed the classifier is trained by giving it sampled data with associated categories. Features are extracted from the data by a feature extraction algorithm; this was built using trial and error as there is a lot of ambiguity in identifying key features within a tweet. Finally using these features the classifier determines probabilities of features associated to each category.

2.3.3.3 Optimization

Once the classifier is trained it was given a test sample to determine its working. The test sample consisted of a part of sample data which was not used in training the classifier. Initially the accuracy was low so we had to tweak the feature extraction algorithm until desirable results were achieved.

2.4 Sentiment Analysis

Websites with micro blogs have become a great source of information from various categories. Using this data to gather information about human conscious and thinking has become a big task for many research and business organizations.

Glivia analyzed approximately 10 million tweets related to Brazils presidential elections for the year 2010. They studied the overall relation of hashtags and human sentiments associated to those hashtags and found a positive co-relation i.e. the sentiments of hashtags matches the sentiments of the person tweeting. They further studied that in twitter peoples sentiments are affected by popular tweets and few popular hashtags and twitter users can affect the overall emotions of the population.

Pang has analyzed sentiments of movie reviews by taking sample of movie ratings and using corpus of movie reviews and forming a standardized set. The number of stars for a given movie serves as a basis for formulating positive, negative and neutral reviews.

In our project we have utilized Naïve Bayes classifier to perform sentiment analysis on tweets as it gave the best results according to Alec Go research in 2009 (as discussed before). Max entropy was also utilized to find sentiment rating of tweets but the results weren't satisfactory.

In the first step a cluster of 5000 tweets were downloaded for various categories. These tweets were bounded by the region of Pakistan to help account for regional language differences. Such as Pakistanis tend to write many slang words in English and they tend to use Urdu phrases in their tweets. This sample was then divided into training and test set with a 90 to 10 ratio. By manually reading tweets they were classified as positive,

negative, neutral or other (those which made no sense). These labels served as baseline to train and then test the classifier.

ID	Set	Tweet	Label
1	Training	Islamabad looks beautiful from Monal	Positive
2	Training	Terror attack in Waziristan	Negative
3	Training	Waiting anxiously for release of wonder woman	Neutral
4	Test	Army and government will win war against terror	Positive

Table 1

Classifier was then trained on these training tweets and then tested on test sample. This gave poor results mainly due to the following reasons:

- Too many feature sets
- Over fitting of classifier
- Some features have more influence on tweet sentiment than others

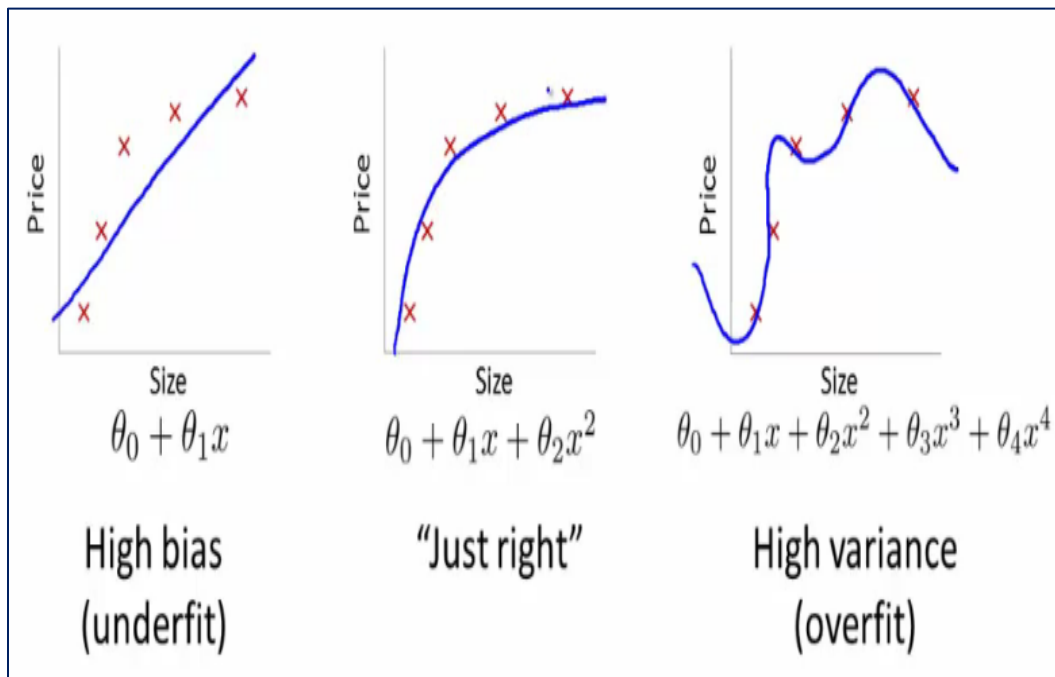


FIGURE 2

To overcome these problems following steps were taken:

Using RE and NLTK common words which have no influence on sentiments were removed, this included prepositions etc. In our first try these words acted as features and our problem became very complex.

Each word in tweet was converted to its root form i.e. running was converted to run and faster to fast etc. This helped simplify our feature set and improved the accuracy of comparisons.

High frequency Urdu words such as those associated to defense, crime and other issues were extracted. From these words a feature set was formed. Using similar approach but with the help of NLTK corpus feature set was formed for English words in root form was formulated as well.

Finally some words which are prone to be associated to basic type of sentiments were given heavier weights as compared to words which conveyed lesser amount of emotions. Such as murder, terrorism, success etc all show either very positive or very negative emotions. Thus their presence in a tweet dictates the type of sentiments the user had.

The classifier was once again formulated and tested using this approach and satisfactory results were achieved.

The results of analysis fit into three main categories i.e. Happy, Sad and Neutral. These results were then displayed in a suitable format to aid users in gathering information quickly and give a general sentiment rating.

The age old saying “one mans trash is another mans treasure” holds true for sentiments.

Some controversial topics can represent both positive and negative sentiments for humans belonging to different communities. Such as something positive for one political party may be a drawback for members of another party. Based on this for future we intend to utilize fuzzy logic for tweet classification to represent both positive and negative emotions of all affiliated parties.

Finally these sentiments were displayed using emoticons to give a quick glance.



2.5 Web Architecture

The website was developed keeping in mind the end users expectations. A very elegant and user friendly page was designed keeping majority of the canvas empty for a central map which would display results of the users query on Google maps.

Tabs were included on the top to help the user select from an array of pre-calculated results to ease the burden on search engine and enhance performance. On the right a list was included which displays information held in tweets. Users can click on individual *emojis* to see the tweet associated to it.

The users had the option of entering key word and the search engine would show tweets associated to it by filtering based on previously determined classifier.

Finally as our application was working on runtime we included bootstrap so the website would continuously refresh on its own and display tweets as they came in.

2.6 Integration

The project consists of two parts, on the backend python scripts and programs prepare data as discussed above and on the front end it consists of an interactive website which displays data and helps user query information according to their will. These parts were integrated by data manipulation and storage. Data was stored in files and they were then read to display it on website. Meanwhile intermediate programs were written which continuously updated and enhanced given files to keep only required tweets and for cleaning purposes to keep the website up-to date

2.7 Visualization

One of the major goals of this project was to enhance visual appearance of tweets and convey spatial information. To do this the concepts of maps were utilized. Tweets were displayed using their co-ordinates on Google maps.

Furthermore the sentiment analysis of tweets was displayed using emoticons which helped give users an instant view of what the user was feeling about a particular topic.

2.7.1 Real-time Mapping

The biggest aspect of this project was to have it running in real-time. This vastly increases the uses of this project as real time analysis can be performed. To do this task program was made which would run autonomously and website was linked to the files. This meant as soon as users made tweets they were downloaded and uploaded on our

website thus keeping it up to date. algorithms were developed to keep information relevant and updated thus keeping front end clean and avoiding possibility of clusters and out dated data.

RESULTS AND DISCUSSIONS

3.1 Working

We were able to achieve our objectives and goals. A real time system was created which downloaded tweets and displayed them on website. The information displayed on the website had sentiment analysis performed on it and gave users an overall abstract of human emotions about particular topics.

The website contains different categories which display tweets only regarding those categories to help user stay up-to date on trending topics.

Finally users have the ability to search for topics that they like and want to get information about and the program will display their wanted tweets thus giving them the power to find topics of their interest.

3.1.1 Connections and Issues

There were some issues with the internet connection as we were connected to online resources, both for extracting location and downloading tweets, thus we required a stable internet connection. Sadly this was not always the case as high ping, or proxy controls on NUST internet stopped the program.

3.2 Trends

The program performed much better than expected on capturing hot trends in Pakistan. It was able to find new trending topics using only the base repository and the classifier was able to distinguish topics which were trending for a long time compared to those which only lasted a few days.

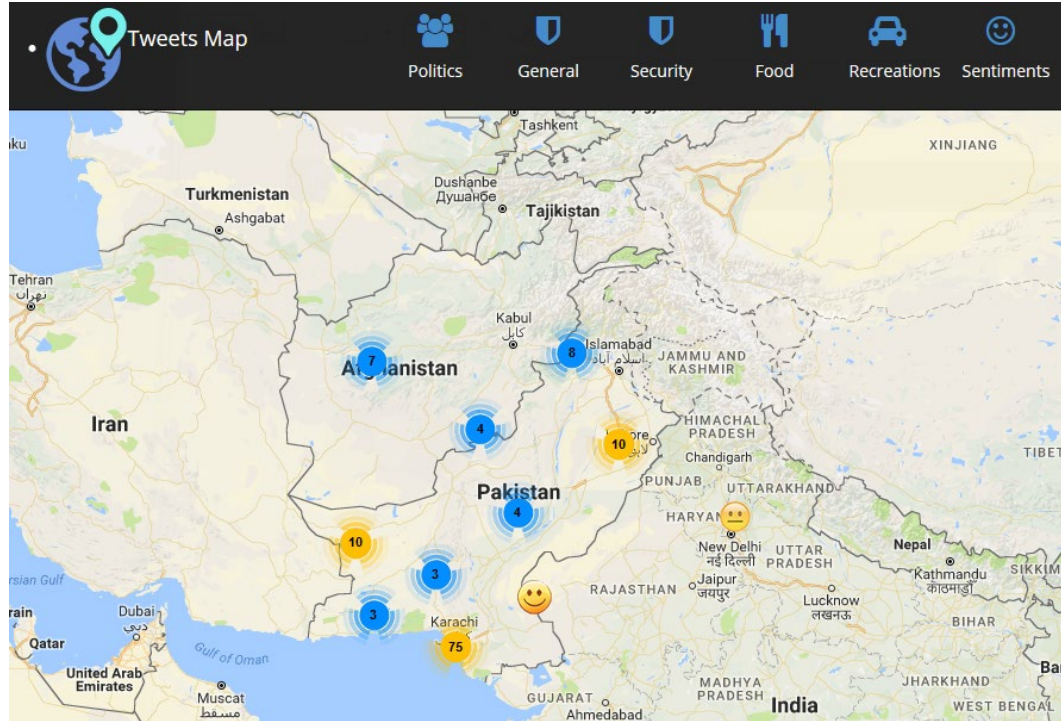
3.2.1 Visual Analysis

Visual analysis was satisfactory, the desired aims were met but information displayed can be greatly enhanced by incorporating 3D display.

Website Front



General Tab




Sentiments

The use of emoticons to display sentiments worked wonderfully, it gave a general picture of what people were feeling about a particular topic. Still there are some issues regarding controversial topics. For future fuzzy logic can be incorporated to help display both sides of the picture.

Sentiments Tab

Politics General Security Food Recreations Sentiments HOME MAPPING



Pakistan

Search sentimental Analysis:
Search..

RT @TaniyaMuniir: Arresting Faisal Ranjha 3rd rate patwari coz he tweeted Ch-tyapa shows #Pakistan establishment is bigger Ch-tya. #BringBa\u20262026
Sat May 20 15:53:48 +0000 2017

@AmirMateen2
@sarwatvalim @KismatZimri
This is a worth procuring to Rauf Klasr\u20262026 for Pakistan--When treason & corruption is protected by

CHAPTER 4

CONCLUSION AND RECCOMENDATION

The project fulfilled its intended purpose and can prove to be a very resourceful utility for young and upcoming tech industries which gain notoriety from social media. Considering its autonomy and ability to refresh and stay up to date on its own it will require very little maintenance.

We recommend anyone interested on improving it to incorporate other social media platforms such as Facebook to enhance its scope and capability. The neural network needs to be more adaptable and flexible to cater for *the human element* of forming associations where they are not apparently present. Finally sentiment analysis is a very vast and changing domain it needs further development, tools like fuzzy logic may be used to show both sides of the emotion and how people on both side of controversial topics feel.

BIBLIOGRAPHY

Mining Sentiments from Tweets Akshat Bakliwal et al International Institute of Information technology, Hyderabad

Detecting Entity-Related Events and Sentiments from Tweets Using Multilingual Resources Alexandra Balahur and Hristo Tanev European Commission Joint Research Centre

Balahur, A.. 2012. The Challenge of Processing Opinions in Online Contents in the Social Web Era Workshop Language Engineering for Online Reputation Management at LREC 2012

Naive Bayes classifiers Kevin P. Murphy

On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes Andrew Y. Ng and Michael I. Jordan

Book “Natural Language Processing with Python 3rd edition”