

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
“JNANA SANGAMA”, BELGAUM – 590014



A Project Report on

“Sentiment Analysis of Tweets”

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

**Bachelor of Engineering
in
Information Science & Engineering**

Submitted by:

| | |
|-----------------------|-------------------|
| AYUSHI SINGH | 1PI13IS029 |
| SATVIK KHETAN | 1PI13IS096 |
| SHUBHAM VATSAL | 1PI13IS105 |

Under the guidance of

Internal Guide
Prof. Raj Alandkar
Department of IS & E,
PESIT



DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING
PES INSTITUTE OF TECHNOLOGY

100 Feet Ring Road, BSK 3rd Stage, Bengaluru – 560085

January 2017 – May 2017

PES INSTITUTE OF TECHNOLOGY

**100 Feet Ring Road, B S K 3rd Stage,
Bengaluru-560085**

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING



CERTIFICATE

This is to certify that the project work entitled “**Sentiment analysis Of Tweets**” carried out by **Ayushi Singh**, bearing USN **1PI13IS029**, **Satvik Khetan**, bearing USN **1PI13IS096**, **Shubham Vatsal**, bearing USN **1PI13IS105**, are bonafide students of **PES INSTITUTE OF TECHNOLOGY**, Bangalore, in partial fulfillment for the award of degree of **BACHELOR OF ENGINEERING IN INFORMATION SCIENCE & ENGINEERING** of **Visvesvaraya Technological University, Belgaum** during the year **2017**. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the above said degree.

Prof. Raj Alandkar
Internal Guide
Department OF ISE
PESIT

Dr. Shylaja S S
HOD
Department OF ISE
PESIT

Dr. K. N. BalasubramanyaMurthy
Principal
PESIT

External Viva

Name of the Examiners

Signature with Date

1. _____

2. _____

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

P.E.S. Institute of Technology

Department of Information Science and Engineering

Bengaluru – 560085



DECLARATION

We, **Ayushi Singh, Satvik Khetan, Shubham Vatsal**, students of Eighth Semester B.E., in the Department of Information Science and Engineering, **P.E.S. Institute of Technology, Bangalore** declare that the project entitled **“Sentiment Analysis Of Tweets”** has been carried out by us and submitted in partial fulfillment of the course requirements for the award of degree of **Bachelor of Engineering in Information Science and Engineering of Visvesvaraya Technological University, Belgaum** during the academic year **2016-17**. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

Name:

USN:

Signature

Acknowledgement

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend my sincere thanks to all of them.

We are highly indebted to **Prof. Raj Alandkar**, our project guide for his guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

My thanks and appreciations also goes to our project coordinators for developing the project, dedicating time and for maintaining the team spirit and people who have willingly helped us out with their abilities.

We would like to thank our head of department, **Dr. Shylaja S** for letting us get such a experience and exposure and for her support and encouragement towards our project.

We would like to thank our principal, **Dr. K. N. Balasubramanya Murthy** for providing us such an opportunity to showcase our skills and learn new things and for supporting and encouraging us.

We would like to express our gratitude towards our parents for their kind co-operation and encouragement which help us in completion of this project.

We would like to express my special gratitude and thanks to our department for giving us such an opportunity.

Abstract

Twitter is among one of the most influencing social media platforms for people to express their feelings in the recent past whenever an event happens in any domain for example when elections happen or something new like demonetization, during any natural calamity, product reviews on the release of new products. This social media site has enormous opulent data that is either structured or semi-structured or non-structured and this data can be converted into useful information and processed. This contributes in making the society take data driven decisions. The 140 character tweets have become a powerful tool for customers /users. These tweets carry a lot of useful information like sentiment, engagement, reviews and features of its products and what not.

Twitter has 317 million active users per month, total number of tweets sent per day is 500 million, 100 million daily active users, 1.3 billion registered users are there on twitter. With so many people present on twitter why just we stay limited to the tweets done by those hundreds and thousands of people who we follow. The more the people, the better the statics and analysis will be done. Companies like flipkart, amazon, Nike, Walmart, apple, Samsung etc. use twitter to improvise their service and keep customer happy by having official accounts for responding the tweets made with a reference to them. Twitter is a big platform to do analysis and improvise business for companies of all domains as mentioned earlier.

Every day, every hour, every minute a new topic is trending on twitter and everyone tweeting their opinion/view/review on it. People read a few tweets on their timeline and getting influenced by them and ending up making opinion on those few tweets on their timeline without even doing any homework or knowing the scenario. So, they should be made aware of the whole scenario by doing analysis on twitter data and providing a clear and better picture of how most of the people think for better opinion formation and to stop the influence of little information as no information is also better than incomplete information.

Table of Contents

| | |
|---|----|
| Chapter I: Introduction | 07 |
| Chapter II: Problem Definition..... | 11 |
| Chapter III: Literature survey..... | 12 |
| Chapter IV: Software Requirement Specification..... | 19 |
| 4.1 Document Purpose..... | 19 |
| 4.1.1 Hardware Interface | 19 |
| 4.1.2 Software Interface | 20 |
| 4.1.2.1 Inputs..... | 20 |
| 4.1.2.2 Outputs..... | 20 |
| 4.1.2.3 Operating System..... | 20 |
| 4.1.3 User Interface | 20 |
| 4.1.3.1 Sheet1: Topic Mood Dashboard..... | 20 |
| 4.1.3.2 Sheet2: Ecommerce issue Dashboard..... | 21 |
| 4.1.3.3 Sheet3: Product Comparison Dashboard..... | 21 |
| 4.1.3.4 Error Notification..... | 21 |
| 4.1.3.5 Retrieving Inputs..... | 21 |
| 4.1.3.6 Real Time Processing..... | 22 |
| 4.1.3.7 Sentiment Analysis..... | 22 |
| 4.1.3.8 Output..... | 22 |
| 4.1.4 Non Functional Requirements..... | 22 |
| 4.1.4.1 System Resource Consumption..... | 22 |
| 4.1.4.2 Safety and Security Requirements..... | 23 |
| 4.1.4.3 Software Quality Attributes..... | 23 |
| 4.1.4.4 Gantt Chart..... | 24 |

| | |
|--|----|
| Chapter V: System Design..... | 25 |
| 5.1 Data Flow Diagram..... | 25 |
| 5.2 Classification Flow Diagram..... | 28 |
| Chapter VI: Implementation..... | 30 |
| Chapter VII: Results and Discussion..... | 39 |
| Chapter VIII: Conclusion..... | 46 |
| Chapter IX: Further Enhancements..... | 47 |
| Chapter X: Bibliography..... | 48 |

CHAPTER I:

INTRODUCTION

We are a part of a society where using internet, using applications that involves communication, etc. has become an essential part of our daily lives and it is growing on a very fast pace because of which many companies are trying to manage this large amount of data to some useful and relevant knowledge to know about people's views and reaction towards their product or domain. Online social networking platforms such as twitter have their large-scale database of user generated content which inturn gives them an opportunity to learn an insight about the emotional nature of a nation and the global community too.

Such great amount of unstructured data can't be manually analyzed, but sites like twitter allow users to contribute, modify and grade the content as well as to express their personal views about any topic. Twitter allows analysis of tweets and have shown that based on several fields, patterns of positive and negative tweets can be observed.

Twitter has become such a influencing platform that be it a company like Nike, flipkart or any big personality like celebrities, politicians etc. everyone stays responsive to the tweets which includes them cause no one wants a bad image or review because people get influenced easily. Companies like Walmart, general motors, Nike, Flipkart, Apple, Samsung etc. have created an official account to respond to their customers.

Doing analysis on twitter data will help companies to understand their audience, to see which content/product resonates through your audience, the engagement, the engagement rate, the followers, the impression etc.

In this era, people have started using social media platforms at such a high rate that they believe whatever they read, they share whatever they feel and they get influenced quickly. Some people in fact get to know about events not through television or newspaper but through social networking sites and there is no guarantee of who is posting what and hence the false information can also spread at a fast rate, so there comes a need to provide those people with a clear and better picture of the topic that are trending on twitter by doing different type of analysis for better opinion formulation and to lessen the effect of incomplete information/bad influence.

Twitter is used mostly for the following reasons:

- Micro-blogging is a famous yet important feature which twitter provides.
- Send and read messages.
- People use it to express their opinion about different topics so it can be used to know about people's opinions.
- People use it for giving product reviews so it can be used to know how a product is by collecting information from many people.
- Awareness posts/informational posts about a current incident to spread awareness, this can be used to spread awareness in the same region etc.
- Twitters audience ranges from regular users to many celebrities in many fields, company representatives, politicians, etc. Therefore, it is possible to collect data from people of different social backgrounds, different ideologies, different religions, different countries and different interests.
- Twitter contains a large number of posts and it grows every day, the collected dataset can be extremely large.

As the number of people on twitter grows every day, data from these sources can be used in opinion mining and sentiment analysis tasks. Our project is all about providing different type of analysis and displaying the results through multiple perspectives to the user.

Twitter is powerful enough to influence people and to be used as a tool. Rather than just reading tweets from your timeline, why not get an overall review of the topic on twitter which is trending before making an opinion? Our project provides an isometric view of the result obtained from analysis on the topic of interest for better understanding Example: Hour wise, location wise etc. It helps you formulate a more appropriate view of the topic at hand so that the bad influence caused by incomplete information can be avoided and our output through Qlikview dashboard makes it more user friendly.

Sentiment Analysis of Tweets

First we will extract tweets on any trending topic and then we will filter the tweets which are related to the topic at hand then after extraction of tweets we clean tweet and then classify them as positive, negative or neutral and perform a few analysis on the data collected like hash tags which have been used in maximum tweets i.e trending hash tags, showing how much viral that topic was hour wise, location wise etc.

Then we finally make a dashboard containing all the charts and details of the analysis which can be seen from any perspective and filtered out results also can be seen.

Our Objective is to provide a better understanding of the situation at hand.

CHAPTER II: PROBLEM DEFINITION

What?

A Prototype to provide more accurate and summarized insights on several topics through twitter data.

The project focuses on using Twitter, the most popular micro blogging platform, for the task of performing various analysis. The tweets are important for analysis because of the availability of large amount of data already present and large amount of data coming very fast and algorithms that process them must do so under very strict constraints of storage and time.

Why?

To make the society take more data driven decisions and opinions. As the people get influenced easily through a few tweets on their timeline, providing a better and clear picture of any trending topic on twitter for better opinion formulation.

How?

It will be shown how to collect a dataset for sentiment analysis and opinion mining purposes using a streaming API and then perform linguistic analysis of the collected corpus. All public tweets posted on twitter are freely available through a set of APIs provided by Twitter. Corpus is cleaned and tokenized and then Using the dataset, a sentiment classifier, is constructed that can determine positive, negative and neutral sentiments.

Using a reporting tool, the results of the visualization of different analysis are enhanced and the results are shown from different perspectives as the output.

CHAPTER III: LITERATURE SURVEY

Data Mining

Some time ago information were not promptly accessible. As information turned out to be more bounteous in the wake of investigating diverse API's, nonetheless, constraints in computational abilities kept the down to earth use of scientific models and to a great degree expansive informational collections brought about framework crash. At present, are information accessible for investigation as well as computational assets can bolster an assortment of refined strategies. Subsequently, announcing devices are presently being utilized for the twitter information investigation result representation.

The bottleneck in information investigation is currently bringing up the most suitable issues for some spaces on twitter and utilizing legitimate information and examination methods to get applicable responses for different areas on twitter.

Information mining is the way towards choosing, investigating and displaying a lot of information. This procedure has turned into an inexorably inescapable movement in every aspect of research and examination.

Information mining has brought about the disclosure of helpful concealed examples from gigantic informational collections. Information mining issues are frequently unraveled utilizing diverse methodologies from both PC sciences, for example, multi-dimensional databases, machine adapting, delicate registering and information perception and insights, including theory testing, cleaning, bunching, grouping, and relapse systems.

Twitter

Twitter is an online news and person to person communication benefit where people can post messages, "tweets," limited to 140 characters along with image. Any person can post tweets, but it is accessible to any individual who visits twitter. Clients get to Twitter through its site

interface, SMS or a cell phone application. Twitter Inc. is situated in San Francisco, California, Joined States, and has more than 25 workplaces around the globe.

In 2013, it was one of the ten most-visited sites and has been portrayed as "the SMS of the Web".

During the USA elections in 2016, Twitter became the biggest flood of news coming from all around the globe with the no of tweets going up to 40 million on this topic

Tweets

Tweets are simple text messages or images that people all around the globe post on twitter. Users can also 'like' any individual's tweets. Twitter allows users to update or change their profile via their mobile phone's app or either by text messaging. Twitter has also been stated as a very simple and easy to use platform by many well-known authorities.

As a social network, Twitter is centered around the principle of followers. When you choose to follow another Twitter user, that user's tweets or posts appear on your main page. If you follow 30 people, you'll see a mixture of tweets coming to your page: new movie review updates, politics updates, even irrelevant rumours.

According to a research the tweets span these six major categories that tells us how world uses twitter across these six categories.

Sentiment Analysis of Tweets

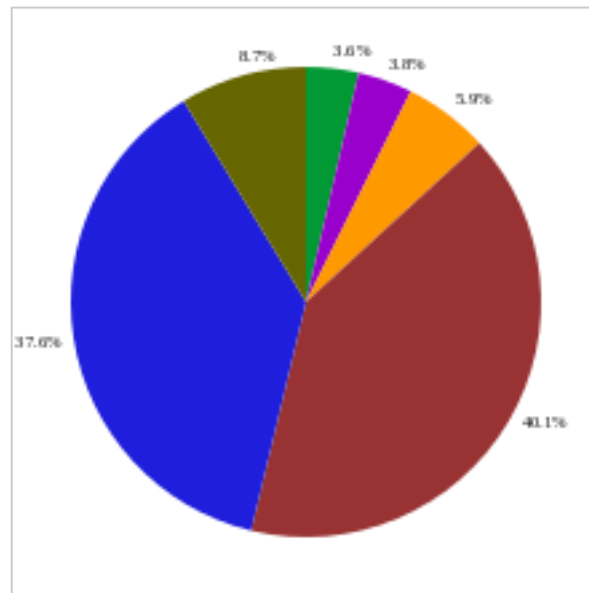


Fig1. Tweet domains.

Content of tweets for each of the six major categories:

News (4.6%)

Spam (4.8%)

Promotions (5.4%)

Irrelevant debates (39.1%)

Conversational (40.6%)

Pass-along value (9.2%)

Another research showcased similar kinds of results where they studied the data on twitter over a span of over 6 months to finally draw these conclusions.

- Irrelevant debate– 40%
- Conversational – 38%
- Pass-along value – 9%

- Promotion – 6%
- Spam – 5%
- News – 5%

Tweet Format

Tweet that is done by different people all around the globe is generally aimed to a trending topic or to a particular person. When they target a trending topic or any other topic they use the respective hashtags in their tweets. For ex: PM Modi wins in UP also #UPElection #BJP , here the person uses the “#” to specify the topic on which he is tweeting. When a user wants to tweet to a particular person or some organization they can use the keyword “@” with their handle name to direct the tweet to the respective personals. For ex: Very bad service @flipkartsupport , by using the @flipkart support the tweets will be directed to the person handling that account.

The tweets are limited to 140 characters which will include emojis, letters, digits, etc. This is very helpful as it restricts the users to post long paragraphs and therefore makes it easy for people to go through these short messages very easily. People can also post images as their tweets.

Data Mining Techniques

Our fundamental approach in this venture has been to order the tweets and after that perform different sorts of examination on the tweets and afterward utilize an envisioning instrument to exhibit our discoveries. We experienced different information mining strategies lastly picked a cross breed classifier of K-Closest Neighbor and Bolster Vector Machine. The three systems have been talked about in detail underneath.

1. K-Nearest Neighbor-This classifier works on the principle of finding the k nearest neighbors, finding max no of neighbors with the same label, assigning the current object the label which is the label of the group with max no of neighbors in it. In sentiment analysis, we can use this classifier to predict the label for a given text or tweet. We use this classifier to look at the labels which can be positive, negative or neutral, of the neighbors and see which label is in majority among the neighbors and assign that label to our current tweet or text.

2. Support Vector Machine- This classifier is widely used in many machine learning algorithms and is a very important part because it offers a clear and accurate decisions amongst the most hearty and precise strategies among all other algorithms. It has a sound hypothetical establishment, requires just twelve cases for preparing, and is uncaring to the quantity of measurements. The SVM can train a model at a very good speed. SVM uses the concept of vector machines and hyperplanes that consists of a boundary so that we can differentiate between positive and negative according to the region in which the current object lies. The accuracy of the model depends upon how well the boundary has been made. The width of the boundary should be of acceptable width so that it will be able to differentiate between the two regions of positive and negative. Once the model is trained with the training dataset and we have the boundary and the 2 regions set then, we can make predictions for a text or a tweet and find out in which region it is supposed to be present and assign the label to the text or the tweet accordingly.

3. Hybrid- A Hybrid classifier is utilized to get the advantages of two existing known classifiers and accordingly helps us in enhancing our precision. For our situation the two existing classifiers are K-nearest neighbor and Bolster Vector Machine. At whatever point another tweet prepares cleaned and is to be arranged, it is first characterized utilizing K-nearest neighbor under the two names in particular "goal" and "subjective". At that point if the tweet is named "objective" then it is pronounced under the mark "unbiased" however in the event that it is named "subjective" then we utilize Bolster Vector Machine classifier to characterize the "subjective" tweet to further "positive" or "negative".

For diving into more insights about Half and half classifiers and diverse courses in which we can examine twitter information, we experienced the accompanying examination papers:

Proceedings of the third international ICWSM conference 2009. Event detection and tracking in social streams

Sentiment Analysis of Tweets

Hassan Sayyadi, Matthew hurts and Alexey Maykov

Event detection over twitter social media streams

Xiangmin Zhou, Lei Chen

Credibility ranking of tweets during high impact events

Aditi Gupta

Data Analysis & Visualization

After collection of tweets using Twitter Streaming API, we clean the tweets and take care of all the stop words and other elements like presence of hashtags and then make our data ready for further analysis. For data visualization, we have made use of a tool named Qlikview. Qlikview is a reporting tool used widely in many companies.

Papers referred:

EventDetectionandsummarization basedon social networks and semanticQuery expansion

K. Sathiyamurthy¹and G. Shanmugavalli²and N. Udayalakshmi³.

Event Detection in Social Streams

Charu C. Aggarwal And KarthikSubbian.

Event detection over twitter social media streams

Xiangmin Zhou and Lei Chen.

Event Identification in Social Media

Hila Becker,CMorNaaman and Luis Gravano.

Online Social Networks EventDetection

A Survey By M´ario Cordeiro¹(B) and Jo˜ao Gama².

Real-Time Classification of Twitter Trends

ArkaitzZubiaag,Damiano Spina, Raquel Mart´inez,V´ıctor Fresno.

Sentiment-Based Event Detection in Twitter

Georgios Paltoglou.

CHAPTER IV:

SYSTEM REQUIREMENT SPECIFICATION

This section is a guide understanding the project - What the product is, its goals and objectives, the target audience and a key to understanding the document conventions.

4.1 Document Purpose

This document specifies the software requirements of sentiment analysis of tweets, developed using Data Mining techniques. This document provides an overview of the said analysis along with the necessary specific and non-specific requirement set.

To be specific, the SRS will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli.

4.1.1 Hardware Interfaces

The application is intended to be a stand-alone, single-user system. The application will run on any computer having Ubuntu/windows installed.

4.1.2 Software Interfaces

4.1.2.1 Inputs

We are concentrating on three types of analysis so for those inputs will be:

- ▶ For a trending topic: The input for the prototype is the tweets related to that topic. It is in the form of an excel file.
- ▶ For ecommerce site analysis: The tweets for @flipkartsupport are used to showcase it.
- ▶ For product comparison analysis: Tweets related to Samsung and Apple are used to provide a user a comparison tab for mobile phones.

4.1.2.2 Outputs

The output will tell us the reaction and how people are responding on a given topic in the form of a dashboard on Qlikview for each of the analysis. If available, historical data will be displayed on the dashboard as well.

4.1.2.3 Operating System

The software will run on the Any operating system which has Ubuntu/windows on it.

4.1.3 User Interfaces

4.1.3.1 SHEET 1: Topic Dashboard

This tab in the dashboard will consist of a few graphs which shows the current trend of the people on Twitter on a given topic. This will be done through several pie charts, line charts, bar graphs, etc. These charts will allow the user to form an opinion around that topic and analyze how it is progressing along with several other analysis.

4.1.3.2 SHEET 2: E-Commerce site issues Dashboard

This graphic will display an overview of the results of each analysis on the e-commerce site under consideration here we have taken @flipkartSupport in the form of a few graphs and charts. This graph should be displayed in a clear and meaningful manner that allows the user to easily interpret the trend of the sentiment toward the topic over all analysis sessions.

4.1.3.3 SHEET 3: Product comparison Dashboard

This sheet contains a few graphs and charts displaying the comparison between any two products for now we have taken apple and Samsung product comparison.

4.1.3.4 Error Notifications

Error notifications is required within the application, presenting the user with appropriate messages such as unable to authorize with twitter, internet connection not present, etc. which describe the error that has taken place. If applicable, error messages should suggest possible solutions to the problem.

4.1.3.5 Retrieving Input

The software will receive three inputs: keywords that is the hashtag the users wants to look into, Keywords will be entered by the user for each topic. Using the secret keys and the tokens the tweets related to that topic will be retrieved with the Twitter Streaming API and then processed and stored in an excel sheet.

4.1.3.6 Real-Time Processing

The software can take input, process data, and display output in real-time. Using this a user can get and insight on any topic in real time with charts and results varying continuously with the continuously coming stream of data.

4.1.3.7 Sentiment Analysis

Sentiment analysis will be done using the hybrid classifier on the text of the tweet to figure out whether tweet is positive or negative or even neutral in respect to the topic.

4.1.3.8 Output

The product must yield all the investigation done on the information as a straightforward dashboard containing sheets. Likewise, the product may yield a diagram of inclination patterns after some time, and also extra insights relating to a point (normal feeling over all investigation sessions and aggregate number of tweets prepared). This yield ought to be clear and straightforward.

4.1.4 Non-Functional Requirements

4.1.4.1 System Resource Consumption

Resource consumption of this application should not reach an amount that renders the usage for a user. The application should be able to operate in the background if the user wish to utilize other applications.

4.1.4.2 Safety and Security Requirements

- 1) Since no login details or sensitive information is stored by the product, it does not run the risk of any unwanted intrusions.
- 2) Safe and secure authentication with Twitter.
- 3) The PC on which the product resides will have its own security.

4.1.4.3 Software Quality Attributes

- 1) The software is planned to be robust and complete, to attract new users, while also providing an intuitive and usable interface that is free of clutter and easy to use.
- 2) Portability is guaranteed since the product is operating system independent.
- 3) Efficiency – This product is not a resource hog and is not computationally intensive.
- 4) Flexibility – This product is modularized and hence is easily expandable, allowing for quick evolution of the software to adapt to possible situations in the future.
- 5) Usability – The user interface is designed to be concise and user friendly. Users can use the product with minimal or no training. User manuals are provided to help as well.
- 4) Reliability – The product functions always provided there's internet connection.
- 7) Maintainability – Standardized design and implementation documents will be provided to maintain the system.

4.1.4.4 Gantt Chart

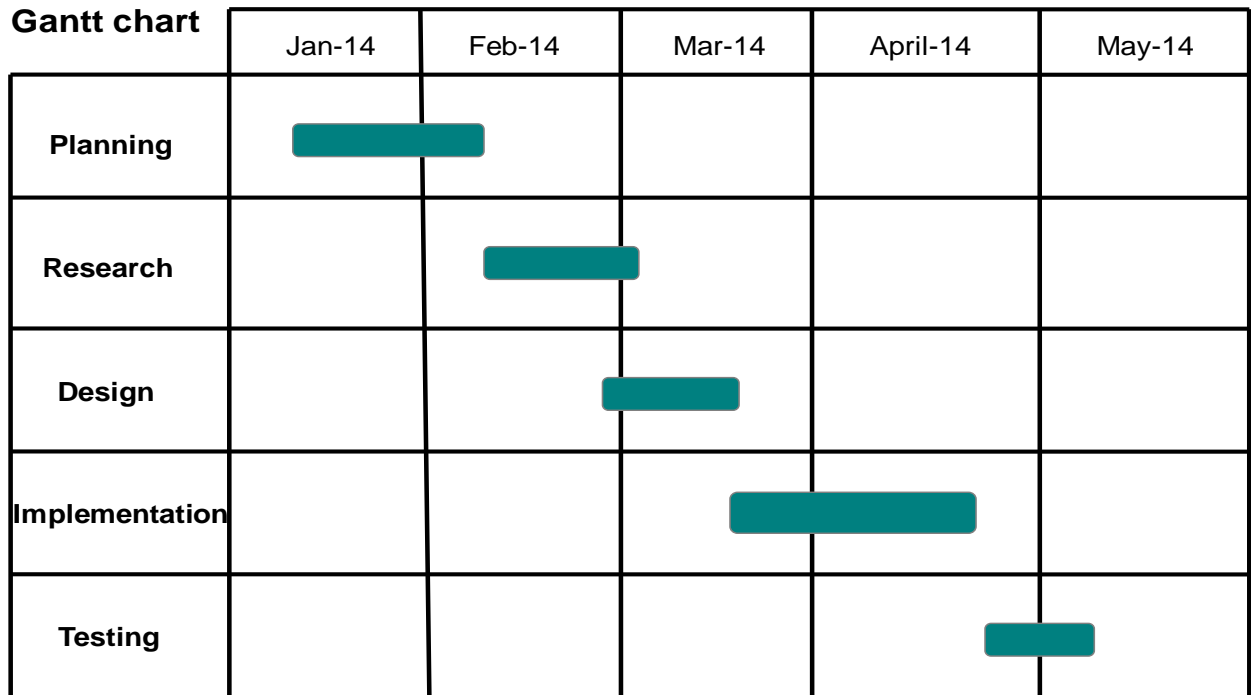


Fig 2. Gantt chart.

This Gantt chart allows us to look at how our project has progressed through the course of 5 months and how much time each of the stages of the project have been utilized in these 5 months.

CHAPTER V: SYSTEM DESIGN

This section describes the design of the system, how the data flows through different modules, and what are the steps followed to process the data collected through twitter to provide the user with an analysis.

5.1 Data Flow Diagram

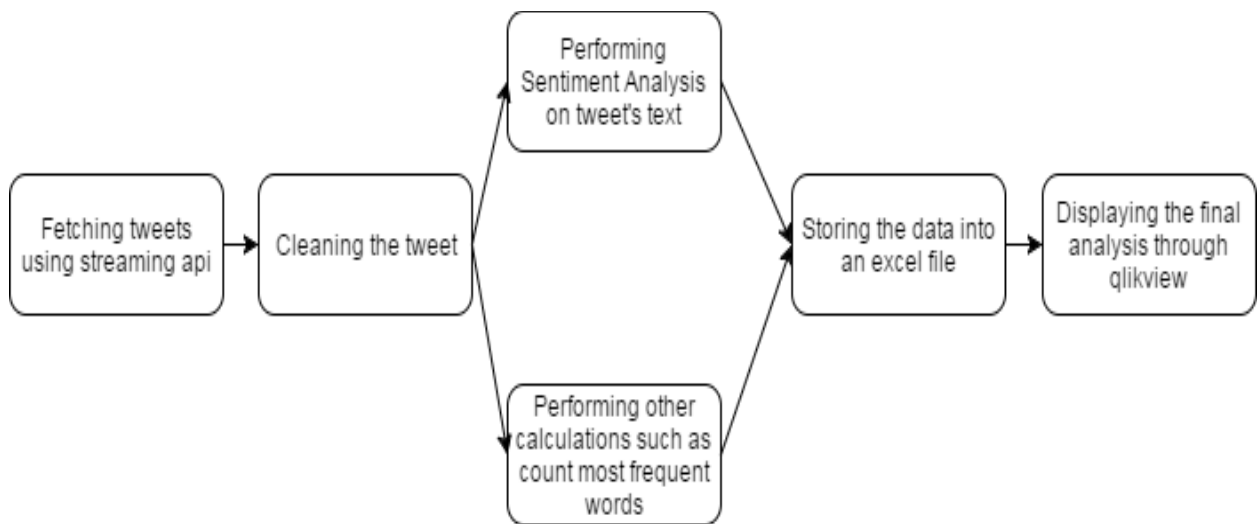


Fig 3 Data Flow Diagram.

Tweepy's streaming api-

- The process starts with fetching the tweets using the streaming api with the help of the tweepy module provided through python.

Input – Authentication Keys to authenticate our account on twitter.

Output – Stream of tweets which are filtered according to user's provided hashtag.

Sentiment Analysis of Tweets

Cleaning Tweet –

- Then several cleaning methods are applied to remove unnecessary text such as links and special characters.

Input – Uncleaned tweet with unnecessary text parts in the tweet.

Output – Text of the tweet with no special characters, hyperlinks, emoticons, etc.

Analysis –

- Then several information regarding the topic and for the whole topic are derived through sentiment classification and counting most frequent related hash tags, etc.

Input – Cleaned text of the tweet to find out its sentiment as well as most frequently occurring hash tags in it.

Output – Sentiment for the tweet and the dictionary of most frequently used hash tags for that topic with keys representing the hash tag and the value representing the count of that hash tag.

Storing the data-

- The complete information is stored in an excel file.

Input – Attributes and different values derived for the tweet

Output – All the data stored in the respective columns in the excel sheet.

Result Visualization-

- This excel file is used as an input in the dashboards created in Qlikview which then shows us different kind of analysis through charts and graphs.

Input – The excel file containing the data for a topic.

Output – Different analysis shown through different charts and diagrams in Qlikview.

5.2 Classification Flow Diagram

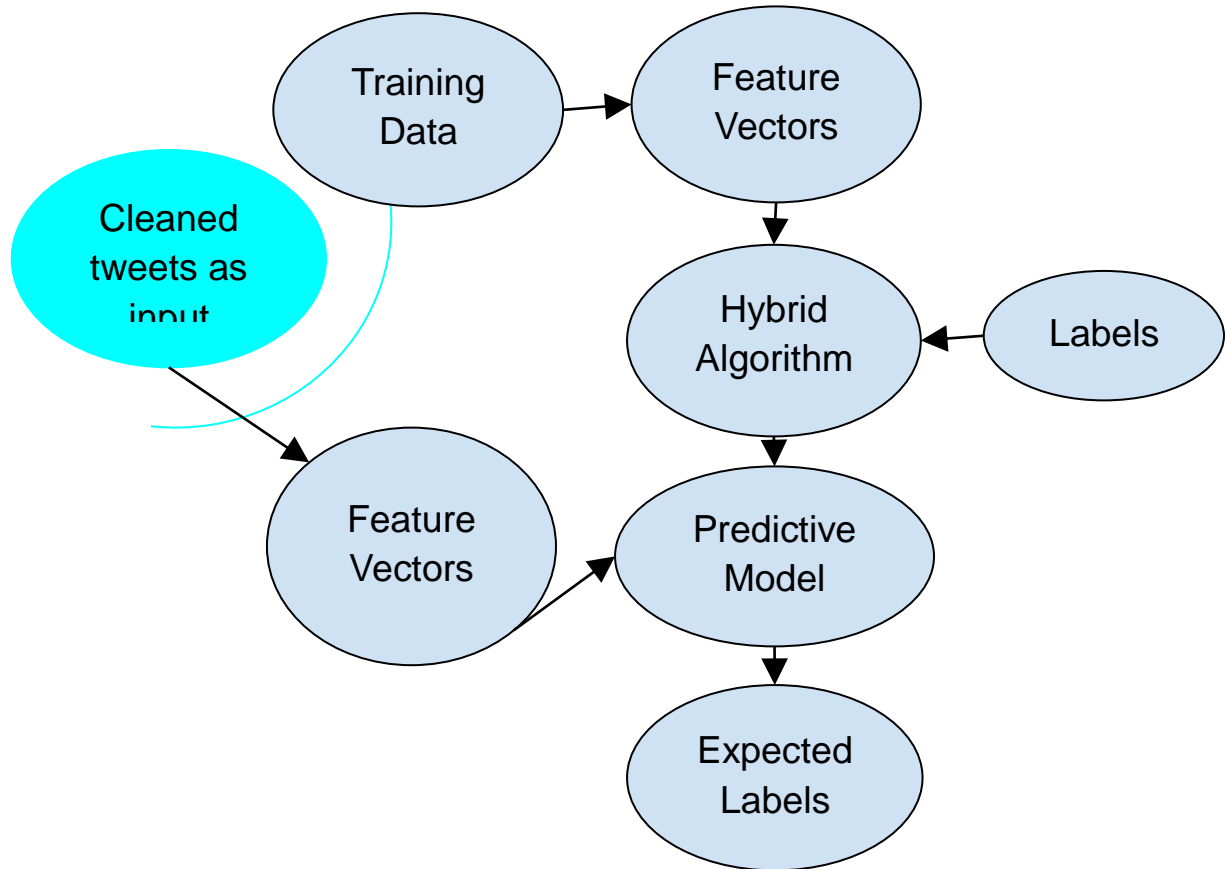


Fig 4 Flow Diagram.

In our work, we are using the hybrid classifier. To use the classifier, we must train it with some training data. These training data is converted to feature vectors which is then sent to the classifier along with the labels.

Providing labels along with the data falls under the category of supervised learning. After the model, has been trained, we can use it to make predictions for data that we send.

This data is in the form of the text which was retrieved from the tweets. After cleaning it, the data is converted to feature vectors so that it can be sent to the model.

Sentiment Analysis of Tweets

After receiving the data, the model predicts the sentiment for the given text and hence we get the sentiment for the tweet that we sent.

CHAPTER VI: IMPLEMENTATION

In our project we used python as the programming language as it provides a library named tweepy that helps us to make calls to the twitter, authorize ourselves and fetch the data according to our needs. We use the streaming api of twitter to fetch the data and process it accordingly to provide the user with various kinds of analysis.

We have a class named StdOutListener() using which we create an object that works like a listener to whenever the data comes on. The class has a function named on_data() which is called everytime the data comes on and an on_error() function which is called whenever there is an error.

```
import tweepy
import json
import csv
import urllib
import hybrid
import re
from datetime import datetime

# Authentication details. To obtain these visit dev.twitter.com
access_token = '327284428-gCvR2imeV7Izmupgl1Y7wYC7Xsq7GBtOg9bw13ap'
access_token_secret = 'NmYEcWCEQmaF9YvYUaIOPxDEGskuiy9tQCxq50dhqtj0F'
consumer_key = 'NEvt6h3PhP0pkMl5LXdfkBvR'
consumer_secret = 'U0pS6I2ZSaYx1KFNUoTcaYhMrHvmBNrKNXBsI4S1BuSfT5eU6D'

accountvar = "#OneNationOneTax" #Search query goes here
outputfilecsv = accountvar+"_stream.csv"
fc = csv.writer(open(outputfilecsv, 'wb'))
fc.writerow(["created_at", "screen_name", "tweet_text", "time_zone", "label"])
#format_time_date="[%b %b %d %H:%M:%S %z %Y]"
# This is the listener, responsible for receiving data
class StdOutListener(tweepy.StreamListener):

    def clean_tweet(self, tweet):
        """
        Utility function to clean tweet text by removing links, special characters
        using simple regex statements.
        """
        return ' '.join(re.sub("(@[A-Za-z0-9]+)|(^0-9A-Za-z \t)|(\w+:\w+\S+)", " ", tweet).split())
```

In `on_data()` function we first decode the json encoded data using the `json.decode()` function from the json library. After decoding we must clean the data to remove special character, links, html characters, etc. Cleaning involves several steps such as removing the HTML characters, removing emoticons using regular expressions, removing stopwords using the list of stopwords from `nltk.corpus` library provided by python.

After cleaning the tweet we call the `myFunction()` function in the hybrid class which will process the text to come up with a value that will represent its sentiment. This function will return an integer value for each sentiment i.e for positive it will return 4, for neutral it will return 2, and for negative it will return 0.

After determining the sentiment of the tweet, we write several tweet's field which are important such as when it was created, location, text of the tweet, name of the person who has written the tweets, etc along with the sentiment that we determined. This process is repeated whenever a new tweet comes up and therefore the excel sheet is populated with the data for a topic that we searched for.

```
def on_data(self, data):
    # Twitter returns data in JSON format - we need to decode it first

    decoded_var = json.loads(data)
    tweet_text=decoded_var['text']
    cleaned_tweet=self.clean_tweet(tweet_text)
    x=int(hybrid.myFunction(cleaned_tweet))
    if(x==4):
        label="positive"
    elif(x==2):
        label="neutral"
    else:
        label="negative"
    time_date=decoded['created_at']
    print(time_date[11:13])
    timeVal=int(time_date[11:13])
    print('%s @%s: %s %s\n' %
    (decoded['created_at'],decoded['user']['screen_name'],decoded['text'].encode('ascii',
    'ignore'),decoded['user']['time_zone'],label,timeVal))
```


Sentiment Analysis of Tweets

```
fc.writerow([decoded['created_at'],decoded['user']['screen_name'],decoded['text'].encode('ascii',
'ignore'),decoded['user']['time_zone'],label,timeVal])

    #got media_url - means add it to the output

    #print "
    return True

def on_error(self, status):
    print(status)

if __name__ == '__main__':
    listener = StdOutListener()
    auth_object = tweepy.OAuthHandler(consumer_key_one, consumer_secret_second)
    auth_object.set_access_token(access_token_one, access_token_secret_second)

    stream_object = tweepy.Stream(auth_object, listener)
    stream_object.filter(track=[accountvar])
```

The following code is responsible for getting the most frequently used hashtags in the tweets collected for a topic.

It contains regular expressions to filter out only hashtags from the tweet's text. The counter will keep a count of the hashtags and finally we select top 10 hashtags according to their frequency and store it in an excel file.

After importing the necessary libraries and initializing lists that are required, we use this code to get us 10 most commonly used words.

```
emoticon = re.compile(r'^'+emoticons_str_list+'$', re.VERBOSE | re.IGNORECASE)

tokens = re.compile(r'('+'.join(regex_str)+')', re.VERBOSE | re.IGNORECASE)

def check(text, lowercase=False):
    tokens_all = tokens(text)
    if lowercase:
        tokens = [token if emoticon_re.search(token) else token.lower() for token in tokens]

def tweetTokenize(text):
    return tokens.findall(text)
```

Sentiment Analysis of Tweets

```
    return tokens

workbook=xlswriter.Workbook('ecommerce_data_count.xlsx')
worksheet=workbook.add_worksheet()

text_list=[]
total=[]
filename='ecommerce_data.csv'
with open(filename,'r') as csvfile:
    reader=csv.reader(csvfile,delimiter=',')
    for row in reader:
        total.append(row)

for screen_name,tweet_text,time_zone,label,hour,location,issue in (total):
    text_list.append(tweet_text)
    csvfile.close()

for text in (text_list):
    print(text)
    print('\n')

count_all_object=Counter()
for tweet in (text_list):
    terms_all_list = [term for term in preprocess(tweet) if term not in stop]
    count_all_object.update(terms_all)

dict_keys= count_all_object.most_common(10)
row=0
col=0

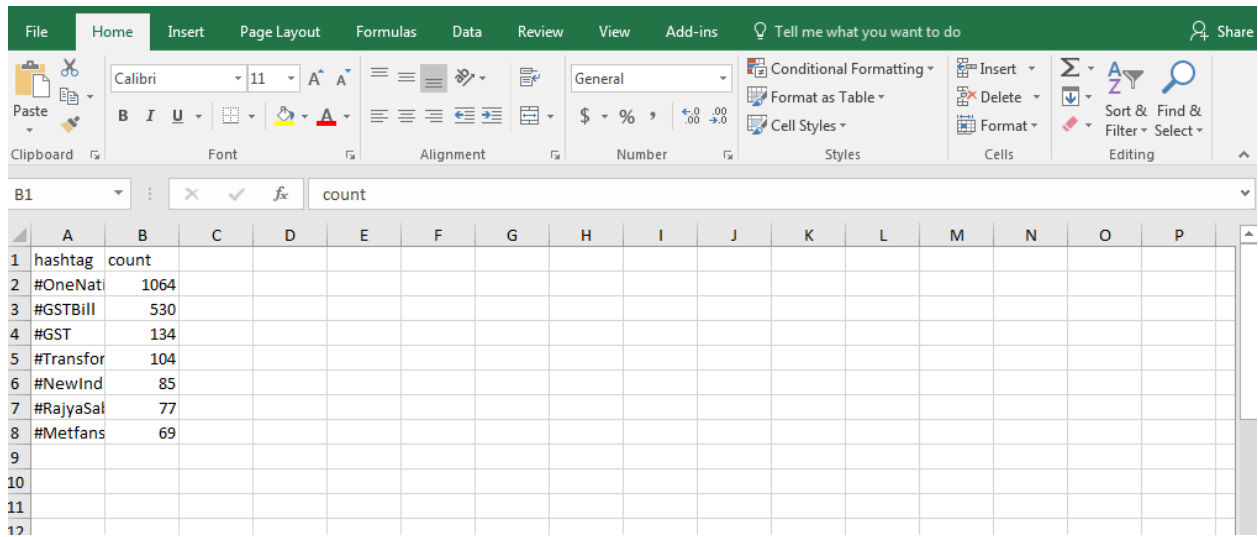
labels, freq=zip(*dict_keys)

for hashtag,count in zip(labels,freq):

    if(hashtag!=':/ and hashtag!="#__"):
        worksheet.write(row,col,hashtag)
        worksheet.write(row,col+1,count)
        row+=1
    workbook.close()
```

After executing this code we were able to get an excel sheet which looks like this:

Sentiment Analysis of Tweets



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|----|-----------|-------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | hashtag | count | | | | | | | | | | | | | | |
| 2 | #OneNati | 1064 | | | | | | | | | | | | | | |
| 3 | #GSTBill | 530 | | | | | | | | | | | | | | |
| 4 | #GST | 134 | | | | | | | | | | | | | | |
| 5 | #Transfor | 104 | | | | | | | | | | | | | | |
| 6 | #NewInd | 85 | | | | | | | | | | | | | | |
| 7 | #RajyaSal | 77 | | | | | | | | | | | | | | |
| 8 | #Metfans | 69 | | | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | | | | | |
| 10 | | | | | | | | | | | | | | | | |
| 11 | | | | | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | | | | | |

Fig 5 Excel Sheet After Code Execution.

The classifier we used is an hybrid classifier which takes advantage of both SVM and kNN to give us an accuracy better than SVM and kNN used separately. The hybrid classifier is as follows:

Importing the necessary libraries and setting the parameters for SVM and KNN classifiers

We import the required libraries

```
import warnings
```

```
import numpy as np
```

```
import features
```

```
import polarity
```

```
import ngramGenerator
```

```
import preprocessing
```

```
warnings.filterwarnings('ignore')
```

```
# User input for model parameters
```

```
N_NEIGHBORS=10 # number of neighbors for KNN
KERNEL_FUNCTION='linear' # kernel function for SVM
C_PARAMETER=0.2
UNIGRAM_SIZE=3000
```

After this we have to initialize the dictionaries such as stop words, slangs etc. and build the unigram vectors using the csv files containing the positive, negative and neutral words. After doing this we have to make sure all the list have no same words and then we have to map the tweets using MapTweet function.

```
print "Initializing dictionnaries"
stopWords = preprocessing.getStopWordList('../resources/stopWords.txt')
slangs = preprocessing.loadSlangs('../resources/internetSlangs.txt')
afinn=polarity.loadAfinn('../resources/afinn.txt')
emoticonDict=features.createEmoticonDictionary("../resources/emoticon.txt")

print "Bulding unigram vector"
positive=ngramGenerator.mostFreqList('../data/used/positive1.csv',UNIGRAM_SIZE) # add as
needed
negative=ngramGenerator.mostFreqList('../data/used/negative1.csv',UNIGRAM_SIZE)
neutral=ngramGenerator.mostFreqList('../data/used/neutral1.csv',UNIGRAM_SIZE)

for w in positive:
    if w in negative+neutral :
        positive.remove(w)

for w in negative:
    if w in positive+neutral :
        negative.remove(w)

for w in neutral:
    if w in negative+positive :
        neutral.remove(w)

# equalize unigrams sizes
m=min([len(positive),len(negative),len(neutral)])

positive=positive[0:m-1]
negative=negative[0:m-1]
neutral=neutral[0:m-1]

def mapTweet(tweet,afinn,emoDict,positive,negative,neutral,slangs):
```

```
    out=[]
    line=preprocessing.processTweet(tweet,stopWords,slangs)
    p=polarity.afinnPolarity(line,afinn)
    out.append(p)
    out.append(float(features.emoticonScore(line,emoDict))) # emo aggregate score be careful to
    modify weights
    out.append(float(len(features.hashtagWords(line))/40)) # number of hashtagged words
    out.append(float(len(line)/140)) # for the length
    out.append(float(features.upperCase(line))) # uppercase existence : 0 or 1
    out.append(float(features.exclamationTest(line)))
    out.append(float(line.count("!")/140))
    out.append(float((features.questionTest(line))))
    out.append(float(line.count("?")/140))
    out.append(float(features.freqCapital(line)))
    u=features.scoreUnigram(line,positive,negative,neutral)
    out.extend(u)
    return out
```

After doing all the processing to train the model using the training dataset we are now ready to make predictions using the model. The function myFunction() allows us to send a text and it return the sentiment of that text by using the hybrid model.

```
def myFunction(text):
    return(predictTwo(text,KNN_MODEL,SVM_MODEL))

# loading training data
X,Y=loadMatrix('../data/used/positive1.csv','../data/used/neutral1.csv','../data/used/negative1.csv',
'4','2','0')
#X,Y=loadMatrix('../data/small_positive_processed.csv','../data/small_neutral_processed.csv','../d
ata/small_negative_processed.csv','4','2','0')

# features standardization
X_scaled=pr.scale(np.array(X))
scaler = pr.StandardScaler().fit(X) # to use later for testing data scaler.transform(X)

# features Normalization
X_normalized = pr.normalize(X_scaled, norm='l2') # l2 norm
normalizer = pr.Normalizer().fit(X_scaled) # as before normalizer.transform([[-1., 1., 0.]]) for
test

X=X_normalized
X=X.tolist()
```

Sentiment Analysis of Tweets

```
# validation step
print "Optimizing "
C=[0.01*i for i in range(1,2)]
N=[i for i in range(10,11)]
ACC=0.0
best_acc=0.0
iter=0
for c in C:
    for n in N:
        print "C parameter : %f, Neighbors %d" %(c,n)
        ACC=validateHybrid(X,Y,n,KERNEL_FUNCTION,c)
        if (ACC > best_acc):
            N_NEIGHBORS=n
            C_PARAMETER=c
best_acc=ACC

print "Model optimized "
print "best c : %f, best n : %d, best accuracy : %f"
%(C_PARAMETER,N_NEIGHBORS,best_acc)

# Building Model
print "Initializing model ..."
KNN_MODEL,SVM_MODEL,s1,n1,s2,n2=buildHybrid(X,Y,N_NEIGHBORS,KERNEL_FUN
CTION,C_PARAMETER)

# test dataset classification, uncomment the next line to perform the test
print "Testing model with test dataset ..."
testFile('./data/test_dataset.csv',KNN_MODEL,SVM_MODEL)
#testFile('./data/small_test_dataset.csv',KNN_MODEL,SVM_MODEL)

print "Model Built . Want to classify a tweet ? ..."
```

When we import this file in any code and use `hybrid.myFunction()` with the text of the tweet, it will provide us with a sentiment value for the tweet which are 4 for positive, 2 for neutral and 0 for negative.

Using these 3 codes allow us to do the analysis on tweets fetched for any given topic or from a particular twitter handle, for example: the ecommerce analysis was done using @flipkartsupport.

Sentiment Analysis of Tweets

In ecommerce analysis, we also had 5 lists using which we were able to determine the issue the customer is facing so that the responsible person will get an idea about it from the analysis and should be able to work according to resolve the issue.

CHAPTER VII: RESULTS AND DISCUSSIONS

Accuracy:

The final analysis is displayed through a dashboard created using Qlikview. It incorporates several number of charts and graphs that helps the user to use these graphs and charts to form an opinion about the data and the topic for which the analysis is being displayed.

We tested the accuracy of all the 3 classifier that are SVM, KNN, and the hybrid classifier over a test data set. The table for the 3 classifiers's accuracy is as followed:

| CLASSIFIER | TEST DATASET | ACCURACY |
|------------|--------------|----------|
| SVM | 200 | 56% |
| KNN | 200 | 54% |
| HYBRID | 200 | 58.6% |

As we can see the highest accuracy is achieved through the hybrid classifier therefore it has been used in our project.

We also tested some tweets manually which a human will know are positive but the classifier predicts out to be negative and vice-versa.

These are some example tweets which were manually tested:

Sentiment Analysis of Tweets

| Tweet Text | Label |
|---|----------|
| RT @ShahnawazBJP: #GSTBill passed in Rajya Sabha. #OneNationOneTax will strengthen the economy. Congrats to PM @narendramodi ji & @arunjait | negative |
| after #OneNationOneTax we should look for #commoncivilcode ... | positive |
| i think by the passes of time people should accept the changes.. | positive |
| #GSTBill is passed in Rajya Sabha today | positive |
| But why exempt J&K? They should work & pay taxes than pelt stones. | positive |
| #OneNationOneTax or one extra tax on citizens? | positive |
| RT @mlkhattar: Congratulating PM Shri @narendramodi ji & FM Shri @arunjaitley ji on getting the landmark #GSTBill passed in #RajyaSabha. #O | negative |
| RT @mlkhattar: Congratulating PM Shri @narendramodi ji & FM Shri @arunjaitley ji on getting the landmark #GSTBill passed in #RajyaSabha. #O | positive |
| RT @kajal_jaihind: @jitu_vaghani @iPankajShukla @narendramodi It is indeed a moment of pride for @BJP4India to witness the passage of | positive |
| Congratulating PM Shri @narendramodi ji & FM Shri @arunjaitley ji on getting the landmark #GSTBill passed in #RajyaSabha. #OneNationOneTax | positive |
| RT @mansukhmandviya: #GSTBill is foundation of #NewIndia. Welcome you all to the worl | positive |

Fig 6 Manually Tested Tweets.

Red labels represent that the tweet was wrongly predicted and the green labels are those tweets which are predicted correctly.

The final analysis are displayed through qlikview and the results are as follows:

Sentiment Analysis of Tweets

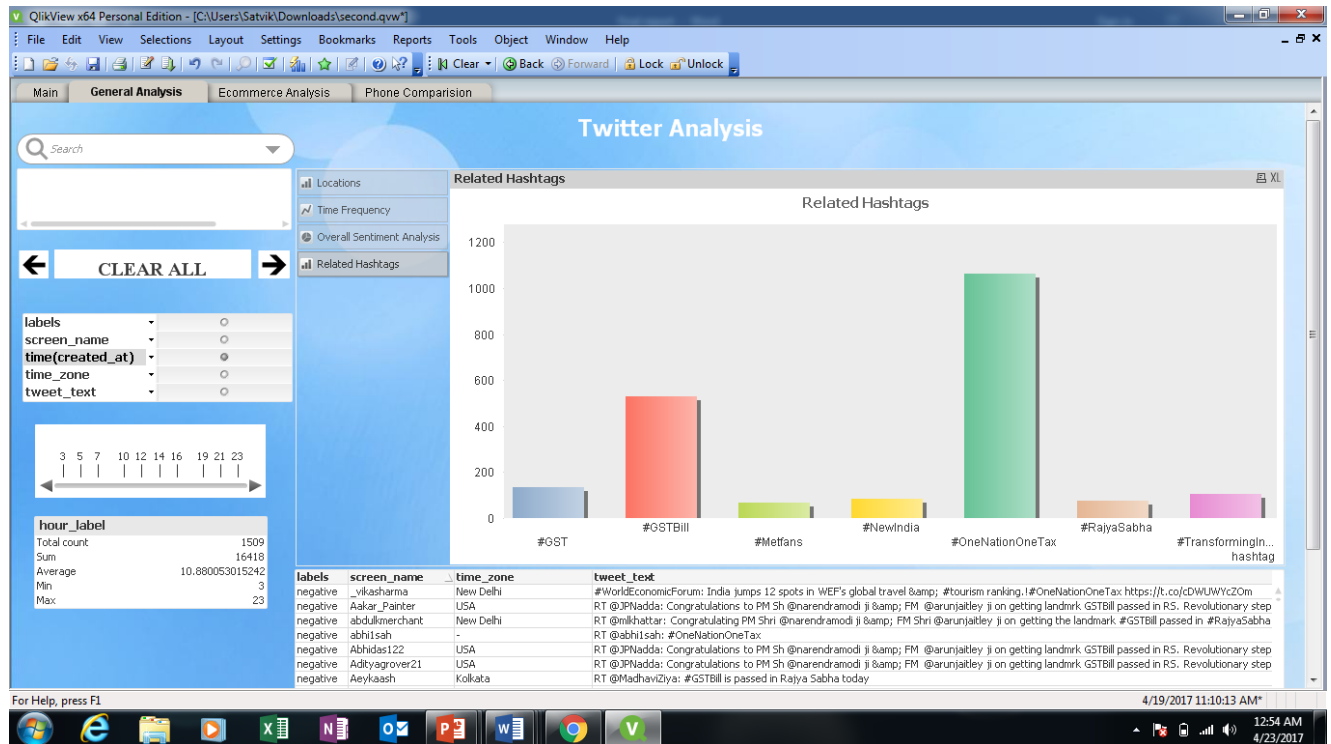


Fig 7 Dashboard Of General Analysis.

This tab shows the frequency of related hashtags for the topic #OneNationOneTax. As we can see #OneNationOneTax is the most frequent with the highest count, second highest is the #GSTBill and so on. This is helpful if we need to see which hashtags are also trending which are related to searched topic.

Sentiment Analysis of Tweets

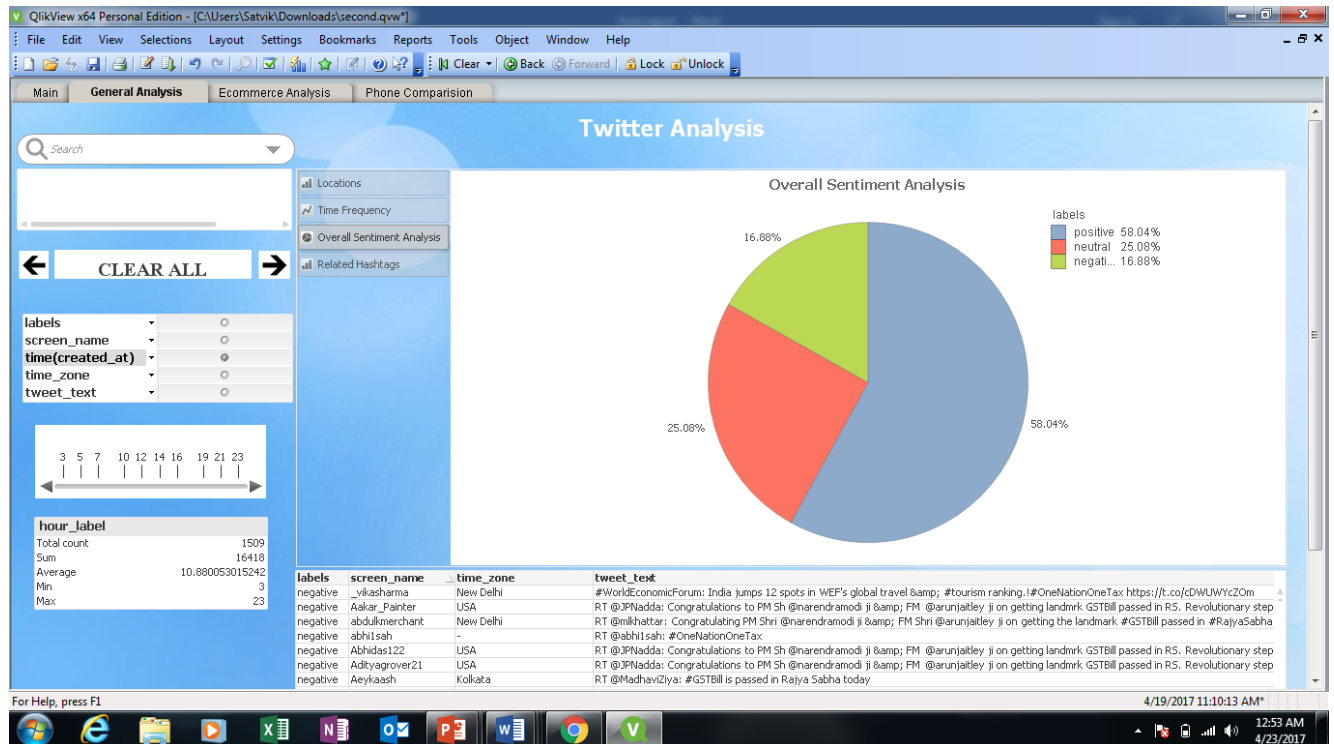


Fig 8 Sentiment Analysis Of Tweets.

This piechart shows the reaction of people about a topic which in this case is #OneNationOneTax. This pie chart shows that 58.04% people are talking positive about #OneNationOneTax on twitter, similarly 25% are neutral and 16% are negative. This piechart can be used for any other topic as well.

Sentiment Analysis of Tweets

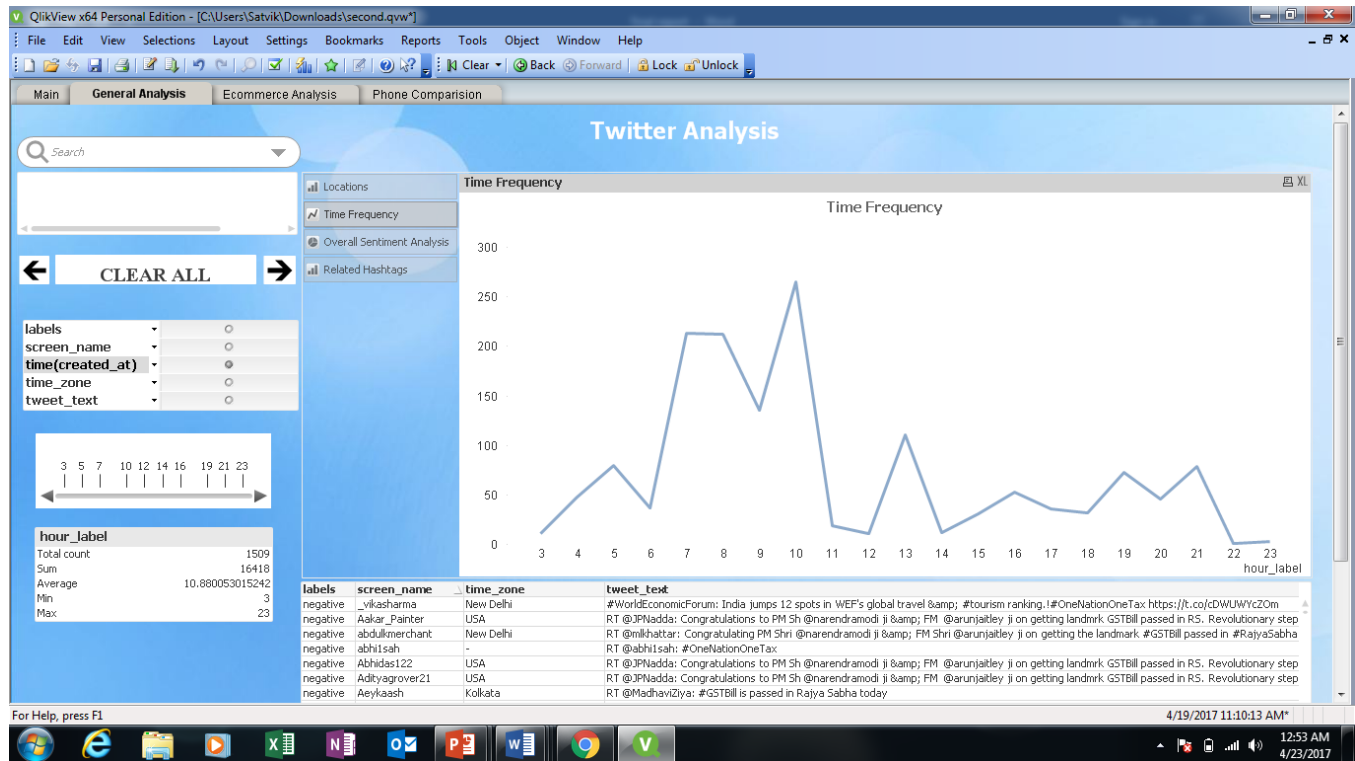


Fig 9 Time Frequency Analysis.

This graph shows the frequency of the tweet at different times of the day. In this graph we can look how tweets on #OneNationOneTax varied across the day in intervals of one hour. This is useful to track the frequency of tweets on other topics as well.

Sentiment Analysis of Tweets

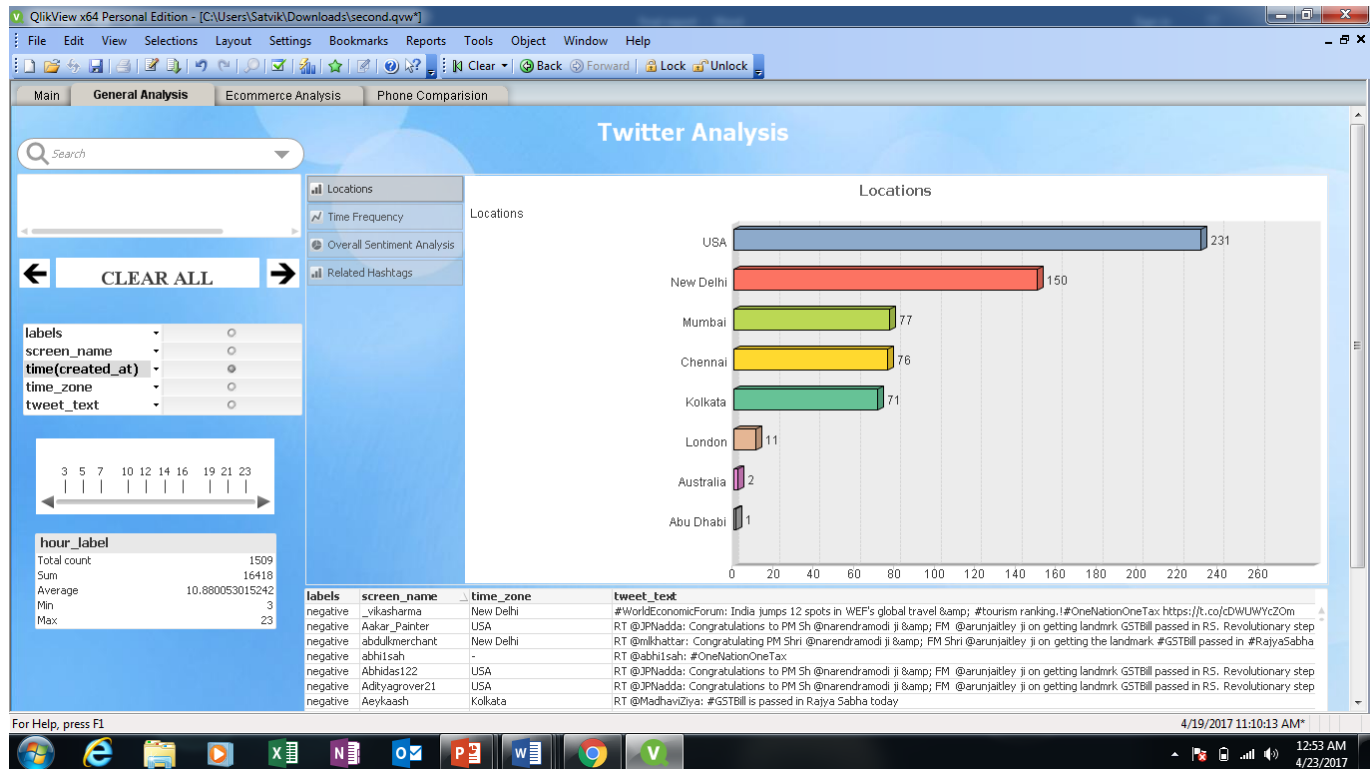


Fig 10 Location Based Analysis Of Tweets.

This bar graph showcase the no of tweets that are generating from different locations around the globe. This kind of analysis are very important during political campaigns as it gives an idea to people of different parties to figure out in which places they are popular and plan their strategies accordingly.

Sentiment Analysis of Tweets

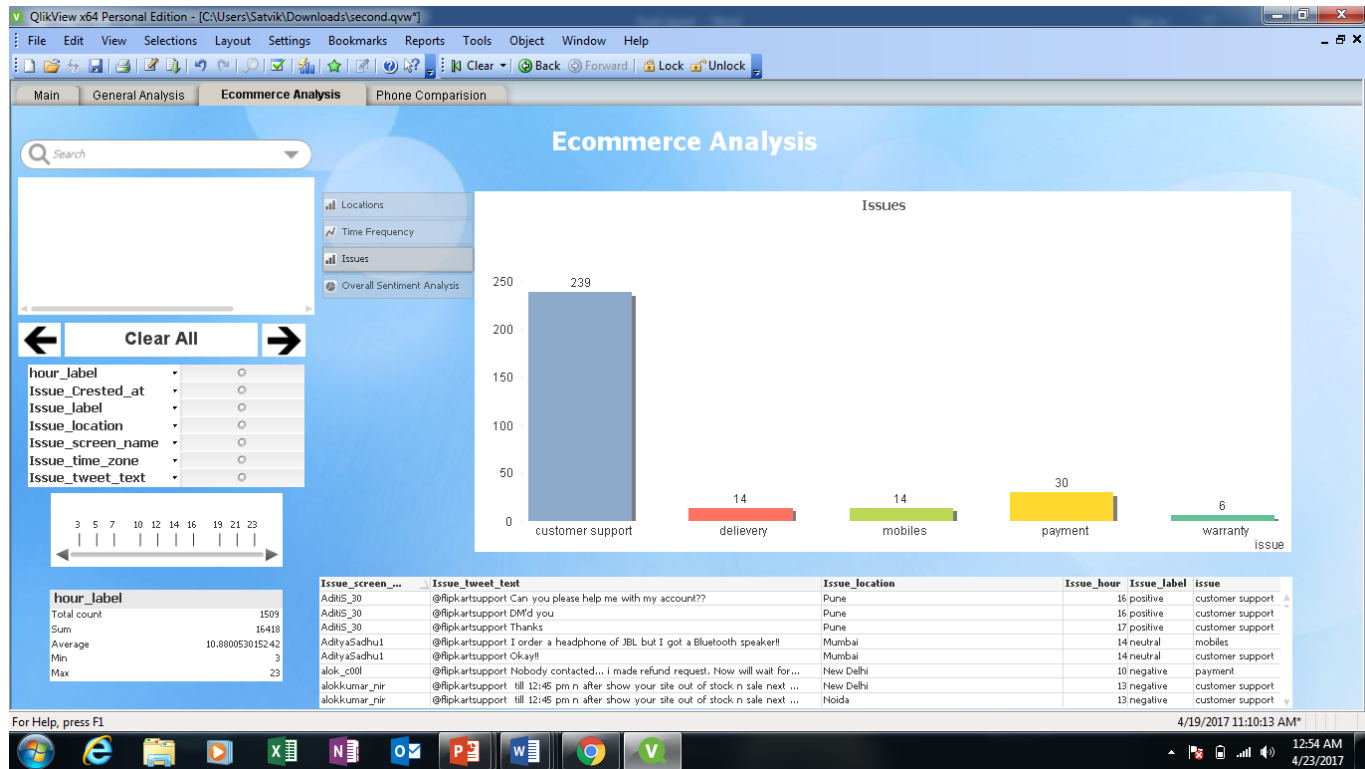


Fig 11 Ecommerce analysis .

This is a tab of ecommerce analysis that shows in which area people are facing most issues. Generally, there are 5 to 6 kinds of issues that people complain about to ecommerce sites ex: customer support, delivery, payment related issue etc. This tab is useful for people that run ecommerce sites as it allows them to keep a track on what people are complaining about so that they can plan accordingly to resolve the issue and focus more on area in which most no of people are having issues.

CHAPTER VIII : CONCLUSION

Application of Data mining in analyzing the tweets is a good method for considering the existing relationships between variables. From our proposed approach, we have shown that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict.

To summarize, the conclusions are:

- Getting an analysis through social media is a very helpful and necessary method for most of the companies.
- In our work, we tried to showcase how some analysis can be withdrawn if the data from twitter is used in a proper way.
- We used a hybrid classifier of SVM and kNN to get better accuracy in predicting the sentiment for a given tweet.
- We used Qlikview to display the analysis as it is the most famous reporting tool in almost all the top companies. It provides a very easy to use interface which allows our application to be very easy to learn and use.
- In our approach, we have tried to keep the front-end user friendly so that a normal person can also use it to gain an insight about some topic he/she is interested in.

CHAPTER IX: FURTHER ENCHANCEMENT

There are some enhancements that can be done in future to increase the productivity and the efficiency of our application. For example:

- Using a complete and latest set of all the tweets around the globe.
- Extending the analysis for Flipkart to other ecommerce sites for example: Amazon, Snapdeal, etc.
- Extending the analysis for reviews on multiple products along with the mobile phones. For example: Reviews for other electronic gadgets, fashion products, etc. can also be included in the range of this analysis.
- Expanding the analysis for other domains as well, ex-political campaigns, government schemes, etc.

Along with this, improvements can also be done with the classifier to provide better accuracy and better efficiency in predicting the sentiment of the tweet.

CHAPTER X:

BIBLIOGRAPHY

- [1] Hassan Sayyadi, Matthew hurts and alexey maykov, Proceedings of the third international ICWSM conference 2009.Event detection and tracking in social streams.
- [2] A SURVEY OF TECHNIQUES FOR EVENT DETECTION IN TWITTER FARZINDAR ATEFEH AND WAEEL KHREICH.
- [3] Bursty event detection from micro blog: a distributed and incremental approach by Jianxin Li*,†, Jianfeng Wen, Zhenying Tai, Richong Zhang and Weiren Yu.
- [4] Detecting Events in Online Social Networks: Definitions, Trends and Challenges by Nikolaos Panagiotou, Ioannis Katakis, and Dimitrios Gunopulos.
- [5] Emerging event detection in social networks with location Sensitivity by Sayan Unankard & Xue Li & Mohamed A. Sharaf.
- [6] Analyzing and Predicting Viral Tweets by Maximilian Jenders,Gjergji Kasneci,Felix Naumann.
- [7] Event Detection and summarization based on social networks and semantic Query expansion by K. Sathiyamurthy¹and G. Shanmugavalli²and N. Udayalakshmi³.
- [8] Event Detection in Social Streams by Charu C. Aggarwal And KarthikSubbian.
- [9] Event detection over twitter social media streams By Xiangmin Zhou and Lei Chen.
- [10] Event Identification in Social Media by Hila Becker,C Mor Naaman and Luis Gravano.
- [11] Online Social Networks Event Detection: A Survey By M´ario Cordeiro¹(B) and Jo˜ao Gama².
- [12] Real-Time Classification of Twitter Trends by Arkaitz Zubiaag,Damiano Spina, Raquel Mart´inez,V´ictor Fresno.
- [13] Sentiment-Based Event Detection in Twitter by Georgios Paltoglou.