TWITTER SENTIMENT ANALYSIS AND VISUALIZATION USING R

A project report submitted

In the partial fulfilment of the requirements of the award of Degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

By

GOUD MANOJ KUMAR (15017T0910)

BANDARI HARSHITHA (15017T0912)

ALI AJEETH AMZATH (15017T0907)

KETHAVATH NARESH (15017T0923)

CHERUKU SHRAVANI (15017T0928)

Under the guidance of

SRI.DR.M.SADANANDAM

Department of Computer Science AND Engineering

UNIVERSITY COLLEGE OF ENGINEERING

KAKATIYA UNIVERSITY

KOTHAGUDEM-507 101

UNIVERSITY COLLEGE OF ENGINEERING

(KAKATIYA UNIVERSITY)

KOTHAGUDEM-507 101

Ph:08744-257123,257125;Fax:08744-25725;email:prinicipal-ku@yahoo.com

CERTIFICATE

This is to certify that the project report entitled **"TWITTER SENTIMENT ANALYSIS AND VISUALIZATION USING R"** is being submitted by

GOUD MANOJ KUMAR-15017T0910

BANDARI HARSHITHA-15017T0912

ALI AJEETH AMZATH-15017T0907

KETHAVATH NARESH-15017T0923

CHERUKU SHRAVANI-15017T0928

in the partial fulfilment for the award of the degree of Bachelor Technology in Computer Science And Engineering to UCEKU Kothagudem is a record of bonafied work carried out under my guidance and supervision.

Project Guide **Head of the Department**

External Examiner

ACKNOWLEDGEMENT

We would like to express our extreme gratitude and sincere thanks to our guide **SRI.DR.M.SADANANDAM , Asst Professor,** for the cooperation, guidance and support.

We express our sincere thanks to **SRI.K.KISHORE KUMAR, Head of the Department, Computer Science AND Engineering,** University college of engineering, Kakatiya University for his encouragement and guidance in carrying out our project.

It gives us immense pleasure in expressing our sincere and deepest sense of gratitude to our **Principal, DR B.SESHA SREENIVASA RAO,** University College of engineering, Kakatiya University, for the facilities made available for the progress and successful completion of our project work.

We would also like to thank our family, friends and also all the teaching and non-teaching staff for supporting and encouraging us during the conduction of the project work.

GOUD MANOJ KUMAR-15017T0910

BANDARI HARSHITHA-15017T0912

ALI AJEETH AMZATH-15017T0907

KETHAVATH NARESH-15017T0923

CHERUKU SHRAVANI-15017T0928

TABLE OF CONTENTS

ABSTRACT

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. There has been lot of work in the field of sentiment analysis of twitter data. This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyse the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. In this paper, we provide a survey and a comparative analyses of existing techniques for opinion mining like machine learning and lexicon-based approaches, together with evaluation metrics. Using various machine learning algorithms like Naive Bayes, Max Entropy, and Support Vector Machine, we provide research on twitter data streams. We have also discussed general challenges and applications of Sentiment Analysis on Twitter.

1. INTRODUCTION

VISUALIZATION

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed. Tables, bar plots, histograms and pie charts can be used for visualization.

TWITTER SENTIMENT ANALYSIS

Twitter is an online news and social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but those who are unregistered can only read them.

Hence Twitter is a public platform with a mine of public opinion of people all over the world and of all age categories. As of October 2016, Twitter has more than 315 million monthly active users. TwitterSentimentAnalysisistheprocessofdeterminingtheemotionaltonebehind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention.

WHY TWITTER SENTIMENT ANALYSIS

The applications for sentiment analysis are endless. It is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics However, it is also practical for use in business analytics and situations in which text needs to be analyzed.

Sentiment analysis is in demand because of its efficiency. Thousands of text documents can be processed for sentiment in seconds, compared to the hours it would take a team of people to manually complete. Because it is so efficient (and accurate – Semantic has 80% accuracy for English content) many businesses are adopting text and sentiment analysis and incorporating it into their processes. Applications: The applications of sentiment analysis are broad and powerful. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

For example, the Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages a head of 2012 presidential election.

The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television adverts

INTRODUCTION TO R LANGUAGE

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- **an effective data handling and storage facility,**
- **a suite of operators for calculations on arrays, in particular matrices,**
- **a large, coherent, integrated collection of intermediate tools for data analysis,**
- **graphical facilities for data analysis and display either on-screen or on hardcopy, and**
- **a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.**

SENTIMENT ANALYSIS

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction.

The words opinion, sentiment, view and belief are used interchangeably but there are differences between them.

- Opinion: A conclusion open to dispute (because different experts have different opinions)

- View: subjective opinion

- Belief: deliberate acceptance and intellectual assent

- Sentiment: opinion representing one's feelings

An example for terminologies for Sentiment Analysis is as given below,

<SENTENCE> = the story of the movie was weak and boring

<OPINION HOLDER> =<author>

<OBJECT> = <movie>

<FEATURE> = <story>

<OPINION >= <weak><boring>

<POLARITY> = <negative>

Sentiment Analysis is a term that include many tasks such as sentiment extraction, sentiment classification, and subjectivity classification, summarization of opinions or opinion spam detection, among others. It aims to analyse people's sentiments, attitudes,

opinions emotions, etc. towards elements such as, products, individuals, topics, organizations, and services.

Mathematically we can represent an opinion as a quintuple (o, f, so, h, t), where o = object; f = feature of the object o; so= orientation or polarity of the opinion on feature f of object o; h = opinion holder; t = time when the opinion is expressed.

Object: An entity which can be a, person, event, product, organization, or topic

Feature: An attribute (or a part) of the object with respect to which evaluation is made.

Opinion orientation or polarity: The orientation of an opinion on a feature f represent whether the opinion is positive, negative or neutral.

Opinion holder*:* The holder of an opinion is the person or organization or an entity that expresses the opinion.

In recent years a lot of work has been done in the field of "Sentiment Analysis on Twitter" by number of researchers. In its early stage it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative only.

Pak and Paroubek(2010) [1] proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, they developed a sentiment classifier based on the multinomial Naive Bayes method that uses features like Ngram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

Parikh and Movassate(2009) [2] implemented two models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets.

They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model.

Go and L.Huang (2009) [3] proposed a solution for sentiment analysis for twitter data by using distant supervision, in which their training data consisted of tweets with emoticons which served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigram were more effective as features.

Barbosa et al.(2010) [4] designed a two phase automatic sentiment analysis method for classifying tweets. They classified tweets as objective or subjective and then in second phase, the subjective tweets were classified as positive or negative. The feature space used included retweets, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS.

Bifet and Frank(2010) [5] used Twitter streaming data provided by Firehouse API , which gave all messages from every user which are publicly available in real-time. They experimented multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They arrived at a conclusion that SGD-based model, when used with an appropriate learning rate was the better than the rest used.

Agarwal et al. (2011)[6] developed a 3-way model for classifying sentiment into positive, negative and neutral classes. They experimented with models such as: unigram model, a feature based model and a tree kernel based model. For tree kernel based model they represented tweets as a tree.The feature based model uses 100 features and the unigram model uses over 10,000 features. They arrived on a conclusion that features which combine prior polarity of words with their parts-of-speech(pos) tags are most important.

A General model for sentiment analysis is as follows,

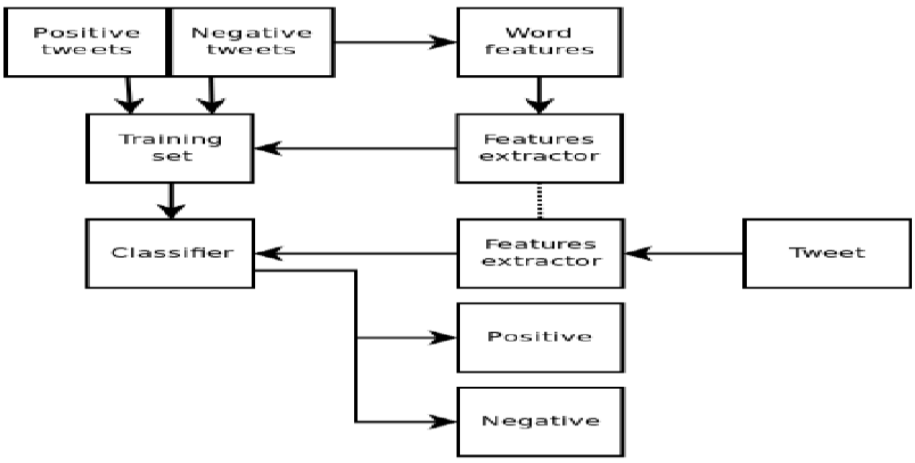


Fig.1. Sentiment Analysis Architecture

Following are the phases required for sentiment analysis of twitter data,

2.1Pre-processing of the datasets

A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labelled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Pre-processing of tweet include following points,

- Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username)
- Correct the spellings; sequence of repeated characters is to be handled
- Replace all the emoticons with their sentiment.
- Remove all punctuations ,symbols, numbers
- Remove Stop Words and Remove Non-English Tweets
- Expand Acronyms(we can use a acronym dictionary)

Table 1. Publicly Available Datasets For Twitter

HASH	Tweets	http://demeter.inf .ed.ac.uk	31,861 Pos tweets 64,850 Neg tweets, 125,859 Neu tweets
EMOT	Tweets and Emoticons	http://twittersenti ment.appspot.co m	230,811 Pos& 150,570 Neg tweets
ISIEVE	Tweets	www.i-sieve.com	1,520 Pos tweets,200 Neg tweets, 2,295 Neu tweets
Columbia univ.dataset	Tweets	Email: apoorv@cs.colum bia.edu	11,875 tweets
Patient dataset	Opinions	http://patientopin ion.org.uk	2000 patient opinions

2.2Feature Extraction

The pre-processed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect are used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram, bigram [18].

Machine learning techniques require representing the key features of text or documents for processing. These key features are considered as feature vectors which are used for the classification

task. Some examples features that have been reported in literature are:

1.Words And Their Frequencies:

Unigrams, bigrams and n-gram models with their frequency counts are considered as features. There has been more research on using word presence rather than frequencies to better describe this feature. Panget al. [23] showed better results by using presence instead of frequencies.

2.Parts Of Speech Tags

Parts of speech like adjectives, adverbs and somegroups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

3.Opinion Words And Phrases

Apart from specific words, phrases and idioms which convey sentiments can be used as features. e.g. cost someone arm and leg.

4.Position Of Terms

The position of a term with in a text can affect on how much the term makes difference in overall sentiment of the text.

5.Negation

Negation is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion.
e.g., I am not happy.

6.Syntax

Syntactic patterns like collocations are used as features to learn subjectivity patterns by many of the researchers.

2.3Training

Supervised learning is an important technique for solving classification problems. Training the classifier makes it easier for future predictions for unknown data.

2.4Classification

2.4.1Naive Bayes:

It is a probabilistic classifier and can learn the pattern of examining a set of documents that has been categorized [9]. It compares the contents with the list of words to classify the documents to their right category or class. Let d be the tweet and c* be a class that is assigned to d, where

$C^{\square} \square \text{ argmac}_c P_{NB}(c \mid d)$

$$P_{NB}(c \mid d) = \frac{(P(c)) \sum_{i=0}^{n} p(f/c)^{ni(d)}}{P(d)}$$

From the above equation, „f" is a „feature", count of feature (fi) is denoted with $n_i(d)$ and is present in d which represents a tweet. Here, m denotes no. of features.

Parameters P(c) and P(f|c) are computed through maximum likelihood estimates, and smoothing is utilized for unseen features. To train and classify using Naïve Bayes Machine Learning technique ,we can use the Python NLTK library .

2.4.2Maximum Entropy

In Maximum Entropy Classifier, no assumptions are taken regarding the relationship in between the features extracted from dataset. This classifier always tries to maximize the entropy of the system by estimating the conditional distribution of the class label.

Maximum entropy even handles overlap feature and is same as logistic regression method which finds the distribution over classes.

The conditional distribution is defined as MaxEnt makes no independence assumptions for its features, unlike Naive Bayes.

The model is represented by the following:

$$P_{ME}(c \mid d,) = \frac{\exp[\sum_i \lambda i f i(c,d)]}{\sum_c \exp[\sum_i \lambda i f i(c,d)]}$$

Where c is the class,d is the tweet and $\lambda_i$is the weight vector.The weight vectors decide the importance of a feature in classification.

2.4.3Support Vector Machine:

Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space[15]. The input data are two sets of vectors of size m each. Then every data which represented as a vector is classified into a class. Nextly we find a margin between the two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it also helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully.

APPROACHES FOR SENTIMENT ANALYSIS

There are mainly two techniques for sentiment analysis for the twitter data:

3.1Machine Learning Approaches

Machine learning based approach uses classification technique to classify text into classes. There are mainly two types of machine learning techniques

3.1.1. Unsupervised learning:

It does not consist of a category and they do not provide with the correct targets at all and therefore rely on clustering.

3.1.2. Supervised learning:

It is based on labeled dataset and thus the labels are provided to the model during the process. These labeled dataset are trained to get meaningful outputs when encountered during decision- making.

The success of both this learning methods is mainly depends on the selection and extraction of the specific set of features used to detect sentiment.

The machine learning approach applicable to sentiment analysis mainly belongs to supervised classification. In a machine learning techniques, two sets of data are needed:

1. Training Set 2. Test Set.

A number of machine learning techniques have been formulated to classify the tweets into classes. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in sentiment analysis.

Machine learning starts with collecting training dataset. Nextly we train a classifier on the training data. Once a supervised classification technique is selected, an important decision to make is to selected, They can tell us how documents are represented.

FUNCTIONALITY AND DESIGN

The process of designing a functional classifier for sentiment analysis can be broken down into five basic categories. They are as follows:

I. Data Acquisition
II. Human Labelling
III. Feature Extraction
IV. Classification

Data Acquistion:

Data in the form of raw tweets is acquired by using the python library “tweetstrem”which provides a package for simple twitter streaming API[26].This API allows two modes of accessing tweets: SampleStream and FilterStream. SampleStream simply delivers a small,random sample of all the tweets streaming at a real time.FilterStream delivers tweet which match a certain criteria.It can filter the delivered tweets according to three criteria:

- Specific keyword(s) to track/search for in the tweets
- Specific Twitter user(s) according to their user -id’s
- Tweets organating from specific location(s)(only for geo-tagged tweets).

A programmer can specify any single one of these filtering criteria or a multiple

Combination of these. But for our purpose we have no such restriction and will thus stick to the SampleStream mode.

Since we wanted to increase the generality of our data,we acquired it in portions at different points of time instead of acquiring all of it at

one go.If we used the latter approach then the generality of the tweets might have been compromised

Since a significant portion of the tweets would be reffering to some certain trending topic and would thus have more or less of the same general mood or sentiment.This phenomenon has been observed when we are going through our sample of acquired tweets.For example the sample acquired near Christamas and New Year's had a significant portion of tweets reffering to these joyous events and were thus of a generally positive sentiment . Sampling our data in portions at different points in time would thus try to minimize this problem. Thus forth,we acquired data at four different points which would be 17th of December 2011,29th of December 2011,19th of january2012 and 8th of February 2012.

A tweet acquiried by this method has a lot of raw

information in which we may or may not find useful for our particular application It comes in the form of the python "dictionary" data type with various key-value pairs. A list of some key-value pairs are given below:

- Whether a tweet has been favourited
- User ID
- Screen name of the user
- Original Text of the tweet
- Presence of hashtages
- Whether it is a re-tweet
- Language under which the twitter user has registered their account
- Geo-tag location of the tweet
- Date and time when the tweet was created

Since this is a lot of information we only filter out the information that we need and discard the rest .For our particular application

we iterate through all the tweets in our sample and save the actual text contents of the tweets in separate file given that

Language of the twitter is user's account is specified to be English. The original text content of the tweet is given under the dictionary key "text "and the language of user's account is given under "Lang".

Since human labelling is an expressive process we further filter out the tweets to be labelled so that we have the greatest amounts of variation in tweets without the loss of generality. The filtering criteria applied are started below:

- Remove Retweets(any tweet which containing the string"RT")
- Remove very short tweets(tweet with length with lessthan 20 characters)
- Remove non-english tweets (by comparing the words of the tweets with a list of 2,000 common English words,tweets with less then 15% of content matching threshold are discarded)
- Remove similar tweets (by comparing every tweet with every other tweet ,tweets with more than 90% of content matching with some other tweet is discarded)

After this filtering roughly 30% of the tweets remain for human labelling on average per sample,which made a total of 10,173 tweets to be labelled.

Human Labelling:

For the purpose of human labelling we made three copies of tweets so that they can be labelled by four indivisualsources.This is done so that we can take average openion of people on the sentiment of the tweet and in this way the noise and inaccuracies in labelling can be minimized Generally speaking the more copies

of labels we can get the better it is ,but we have to keep the cost of labelling in our mind,hence we reached at the reasonable figure of three.

We labelled the tweets in four classes according to sentiments expressed/observed in the tweets :positive,negative,.neutral /objective and ambiguous We gave the following guidelines to our labelers to help them in the labeling process:

- Positive: if the entire tweet has a positive /happy/excited/joyful attitude or if something is mentioned with positive connotations.Also if more than one sentiment is expressed in the tweet but the positive sentiment is more dominant Example *:"4 more years of being in shithole Australia then I move to the USA!:D".*
- Negative:if the entire tweet has a negative/sad/displeased attitude or if something is mentioned with negative connotations.Also if more than one sentiment is expressed in the tweet but the negative tweet is more dominant.Example: *"I want an android now this IPhone is boring:S".*
- Neutral/Obejective: if the creator of tweet express no personal sentiment/opinion in the tweet and merely transmits information.

  Advertisements of products would be labelled under this category .

  Example:ÜS House speaker vows to stop Obama contraceptive rule…
- Ambiguous:if more than one sentiment is expressed in the tweet which are equally potent with no one particular sentiment standing out and becoming more obvious. Also if it is obvious that some personal opinion is being expressed here but due to lack of reference to context it is difficult/impossible to accuirately decipher the sentiment expressed.Example:Ï

Kind of like heroes and don't like it at the same time…"Finally if the contest of the tweet is not apparent from the information available.Example:"That 's exactly how I feel about avengers haha".

- <Blank>:Leave the tweet unlabelled if it belongs to some language other than English so that it is ignored in the training data .

Besides this labelers were instructed to keep personal biases out of labelling and make no assumptions ,i.e. judg the tweet not from any past extra personal information and only from the information provided in the current individual tweet

Once we had labels from four soures our next step was to combine opinions of three people to get and averaged opinion.The way we did this is through majority vote

So for example if a particular tweet had to two labels in agreement ,we would label the overall tweet has such. But if all three labels were different, we labelled the tweet as unable to reach a majority vote". We arrived at the following statistics for each class after going through majority voting.

- Positive :2543 tweets
- Negative:1877 tweets
- Neutral:4543 tweets
- Ambiguous:451 tweets
- Unable to reach majority vote :390 tweets
- Unlabeled non-English tweets:369 tweets

So if we include only those tweets for which we have been able to achieve a positive, negative or neutral majority vote, we are left with 8963 tweets for our training set. Out of these 4543 are objective tweets and 4420 are subjective tweets (sum of positive and negative tweets).

We also calculated the human-human agreement for our tweet labelling task , results of which are as follows:

	Human 1:Human2	Human 2:Human3	Human1:Human3
Strict	58.9%	59.9%	62.5%
Lenient	65.1%	67.1%	73.0%

Table 4 : Human -Human agreement in Tweet Labelling

In the above matrix the “strict” measure of agreement is where all the label assigned by both human beings should match exactly in all cases, while the “lenient “measure is in which if one person marked the tweet as ambiguous” and the other marked it as something else , then these would not count as a disagreement . So in case of the “lenient “measure, the ambiguous class could map to any other class so since the human-human agreement lies in the range of 60-70%(depending upon our definition of agreement),this shows us that sentiment classification is inherently a difficult task even for human beings. We will now look at another table presented by Kim et a1.which shows human-human agreement in case labelling individual adjectives and verbs. [14]

	Adjectives	Verbs
	Human1: Human2	Human1:Human3
Strict	76.19%	62.35%
Lenient	88.96%	85.06%

Table 5: Human-Human agreement in verbs/Adjectives Labelling[6]

Over here the strict measure is when classification is between the three categories of positive, negative and neutral, while the lenient measure the positive and negative classes into one class, so now humans are only classifying between neutral and subjective classes. These results reiterate our initial claim that sentiment analysis is an

inherently difficult task .These e results are higher than our agreement results because this case humans are being asked to label individual words which is an easier task labeling entire tweets

Classification:

Pattern classification is the process through which data is divided into different classes according to some common pattrens which are found in one class which differed to some degree with the pattrens fount in the other classes. the ultimate aim of our project is to design a classifier which accurately classifies tweetsw in the following four sentiment classes :positive,negative,neutral,and ambiguous

There can be twoo kinds of sentiment classifycations in this area: contextual sentiment analysis and general sentiment analysis.

Contextual sentiment analysis deals with classifying specific parts of a tweet according to the context provided . for ex. For the tweet “”4 more years of being in shith hole Australia then I move to the USA:D ” a contextual sentiment classifdier would identify Australia with negative sentiment and usa with a positive sentiument on the other hand general sentiment analysis deals with the general sentiment of the entire text as a whole. Thus

For the tweets menctioned earlier since, there is an overall positive attitude,an accurate general sentiment classifier would identify it as podsitive.

The classification approach generally followed in this domain is the two step approach. First objectively classification is done which deals with classifieng a tweet or a phradse as either objectivr or subjective. After this we perform polarity classification to determine wheather the tweet is positive, negative or both. This was presented by Wilson et a1. and reports enhanced accuracy than a simple one-step approach[16]

We propose a mnovel approach which is slightly different from the aoproach proposed by Wilson et a1[16]. We propose that in first stenpeasch tweet should undergo two classifies : the objectively classifier and the polarity classifier. The former would try to classiofy a tweet between objective and subjective classes, whiloelatter would do so between the positive and negative clasess. We use the shortlisted features for this classifications and use the naïve bayesalgorithenm so nthat after the first step we have two numbers from 0 to 1 representing each tweet .

One of thes numbers is the probability of tweet belonging to objective class and other number is probability of tweet belonging to positive class.since we ccan easily caluculate the two remaining probabilities are subjective or negative by simplewsubstraction by 1,we don’t need those two probabilities.

The most commonly used features in sentiment classification are

- Term presence and their frequency
- Part of speech information
- Negations
- Opinion words and phrases

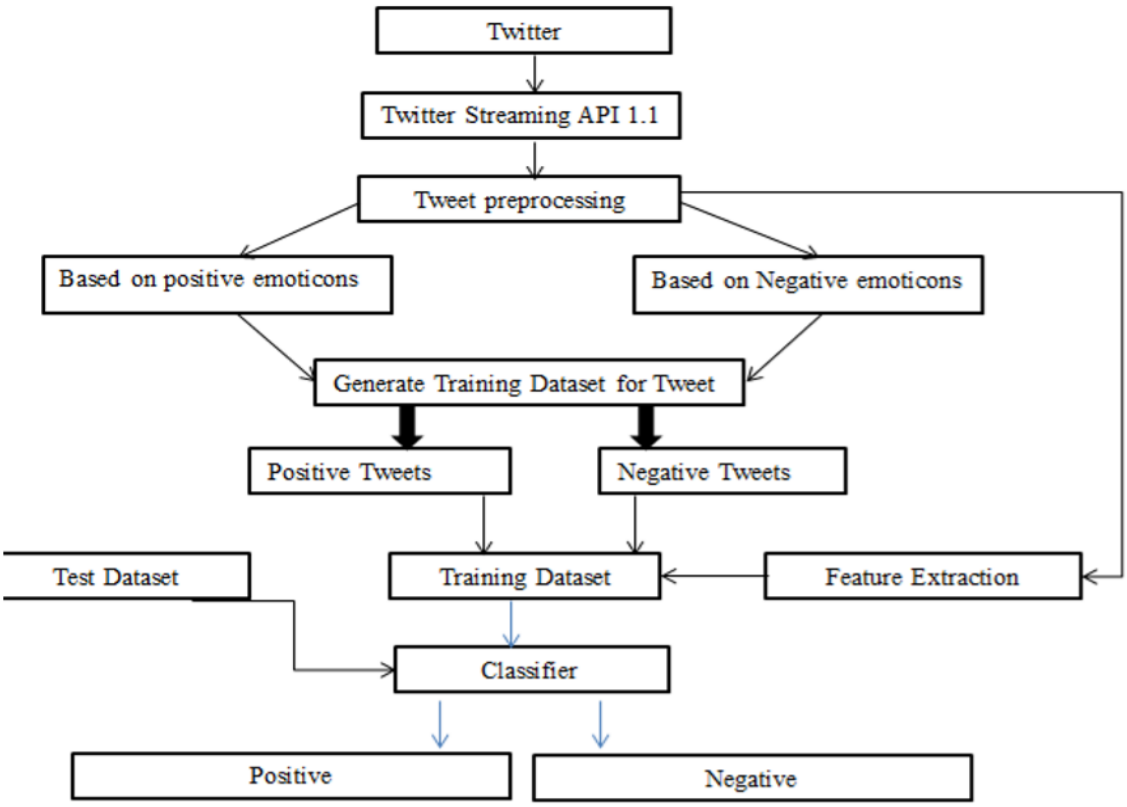


Fig.2 Sentiment Classification Based On Emoticons

With respect to supervised techniques, support vector machines (SVM), Naive Bayes, Maximum Entropy are some of the most common techniques used.

Whereas semi-supervised and unsupervised techniques are proposed when it is not possible to have an initial set of labeled documents/opinions to classify the rest of items

3.2Lexicon-Based Approaches

Lexicon based method [20] uses sentiment dictionary with opinion words and match them with the data to determine polarity. They assigns sentiment scores to the opinion words describing how Positive, Negative and Objective the words contained in the dictionary are. Lexicon-based approaches mainly rely on a sentiment lexicon, i.e., a collection of known and precompiled sentiment terms, phrases and even idioms, developed for traditional genres of communication, such as the Opinion Finder lexicon;

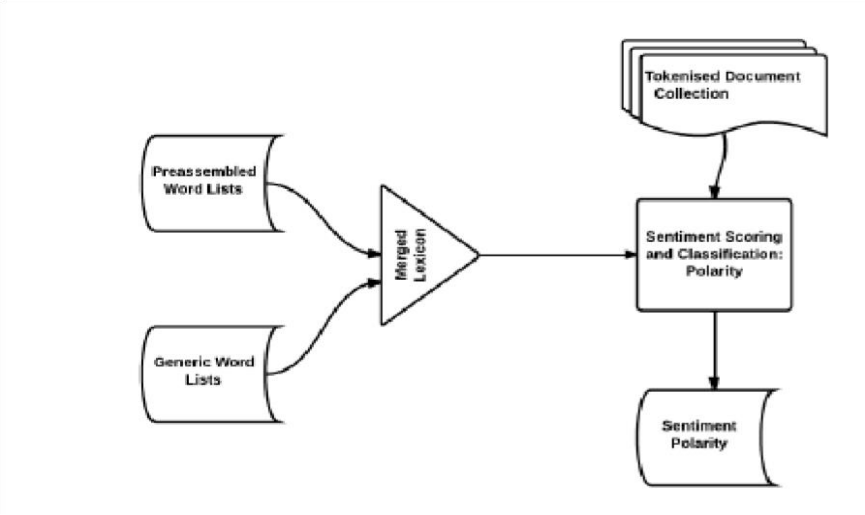


Fig 3.Lexicon-Based Model There are Two sub classifications for this approach:

3.2.1. Dictionary-based:

It is based on the usage of terms (seeds) that are usually collected and annotated manually. This set grows by searching the synonyms and antonyms of a dictionary. An example of that dictionary is WorldNet, which is used to develop a thesaurus called SentiWordNet.

Drawback: Can't deal with domain and context specific orientations.

3.2.2. Corpus-Based:

The corpus-based approach have objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either statistical Techniques:

- Methods based on statistics: Latent Semantic Analysis (LSA).
- Methods based on semantic such as the use of synonyms and antonyms or relationships from thesaurus like WorldNet may also represent an interesting solution.

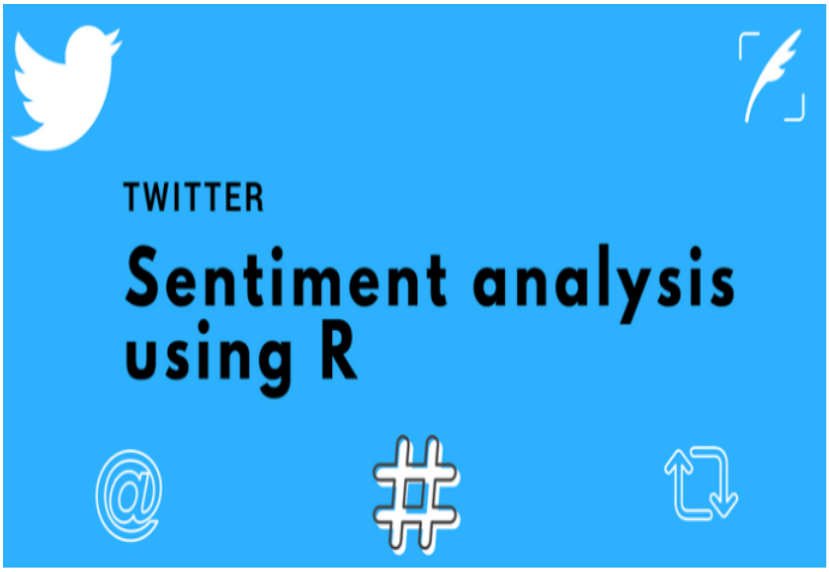
According to the performance measures like precision and recall, we provide a comparative study of existing techniques for opinion mining, including machine learning, lexicon-based approaches, cross domain and cross-lingual approaches, etc.,

OVERVIEW

Tweets are imported using R and the data is cleaned by removing emoticons and URLs. Lexical Analysis is used to predict the sentiment of tweets and subsequently express the opinion graphically through ggplots, histogram, pie chart and tables.

SYSTEM REQUIREMENTS

- Installation of R
- Twitter Authentication to access API



In the past one decade, there has been an exponential surge in the online activity of people across the globe. The volume of posts that are made on the web every second runs into millions. To add to this, the rise of social media platforms has led to flooding to content on the internet.

Social media is not just a platform where people talk to each other, but it has become very vast and serves many more purposes. It has become a medium where people

- **Express their interests.**
- **Share their views.**
- **Share their displeasures.**
- **Compliment companies for good and poor services.**

creating the Twitter app

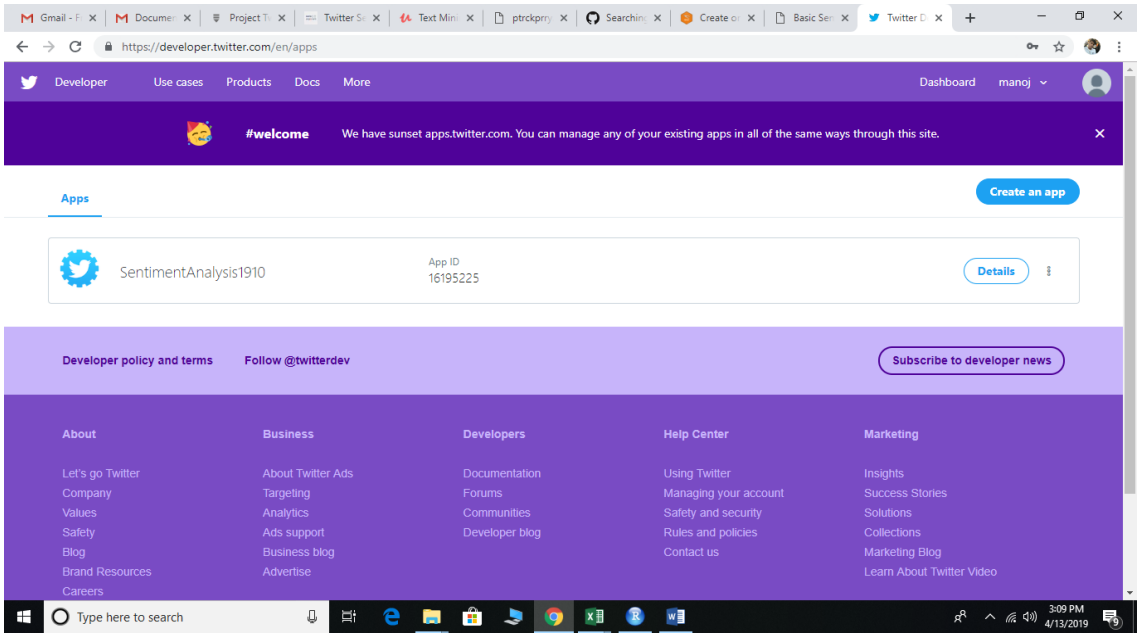
Twitter has made the task of analyzing tweets posted by users easier by developing an API which people can use to extract tweets and underlying metadata.

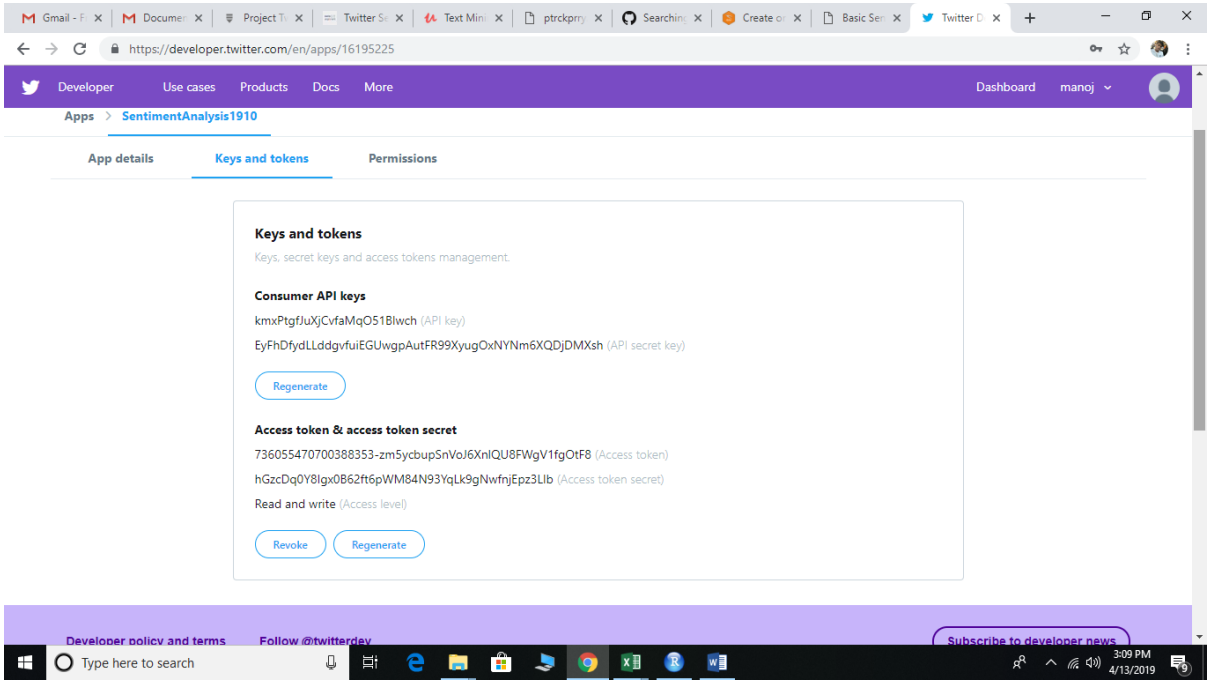
This API helps us extract twitter data in a very structured format which can then be cleaned and processed further for analysis.

To create a Twitter app, you first need to have a Twitter account. Once you have created a Twitter account, visit Twitter's app page and create an application.

Write the basic details such as application name, description along with a website name. You may enter any test website name as well. Once you have entered these details, you will get keys and access tokens. You will get 4 keys and tokens:

1. Consumer Key (API Key)
2. Consumer Secret (API Secret)
3. Access Token
4. Access Token Secret
5. These keys and tokens will be used to extract data from Twitter inR





Challenges in performing sentiment analysis on twitter tweets

Given all the use cases of sentiment analysis, there are a few challenges in analyzing tweets for sentiment analysis. The first one is data quality. The Twitter application helps us in overcoming this problem to an extent.

After basic cleaning of data extracted from the Twitter app, we can use it to generate sentiment score for tweets. The second problem comes in understanding and analyzing slangs used on Twitter.

People have a different way of writing and while posting on Twitter, people are least bothered about the correct spelling of words or they may use a lot of slangs which are not proper English words but are used in informal conversations.

There is a lot of research going on in this area and a lot of people have been able to develop slang dictionaries to understand their meaning. We won't be focusing on this part in this article; we will use the standard dictionaries and packages available in R for sentiment analysis.

The third and the biggest problem in sentiment analysis is decoding sarcasm. Since sentiment analysis works on the semantics of words, it becomes difficult to decode if the post has a sarcasm.

Implementing sentiment analysis application in R

Now, we will try to analyze the sentiments of tweets made by a Twitter handle. We will develop the code in R step by step and see the practical implementation of sentiment analysis in R.

The code is divided into following parts:

1. Extracting tweets using Twitter application
2. Cleaning the tweets for further analysis
3. Getting sentiment score for each tweet
4. Segregating positive and negative tweets

1.Extracting tweets using Twitter application

We will first install the relevent packages that we need.

1.We need package “twitteR”,to extract tweets from twitter

2.”Shiny” is an R package that makes it easy to build interactive web applications straight from R.

----------install.packages(“shiny”)

3.“stringR” package provides a cohesive set of functions designed to make working with strings as easy as possible.

4.”plyr” package-plyr is a R package that makes it simple to split data apart,do stuff to it,and mash it back together.this is common data-manipulation step.

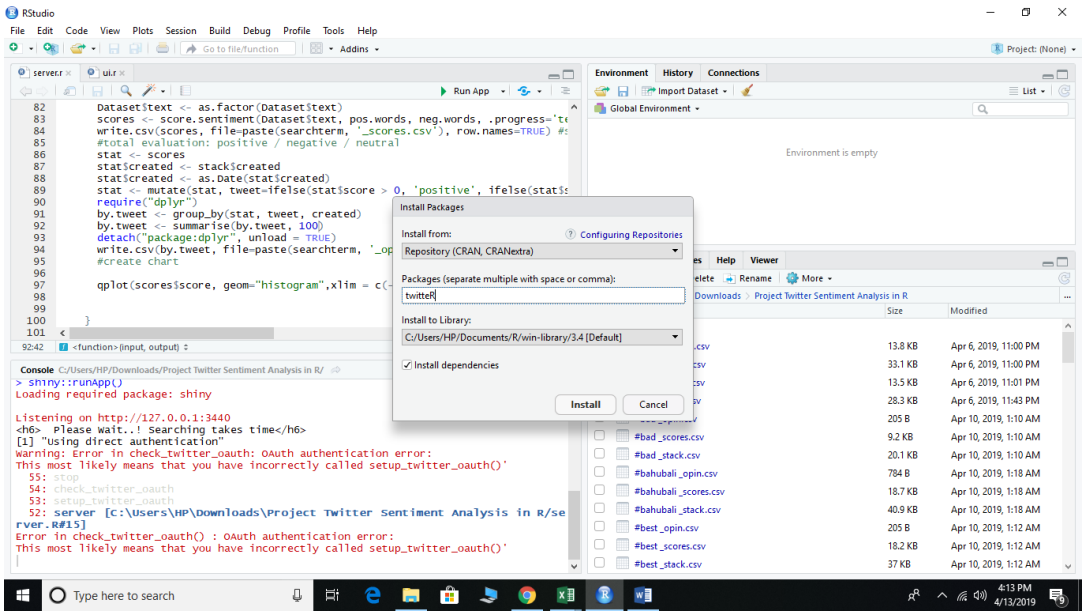
5.”ROAuth” package provides an interface to OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice..here twitter

6.”ggplot2” package allows you to create graphs that represents both univariate and multivariate numerical and categorical data in a straight forward manner.grouping can be represented by color,symbol,size and transparency.

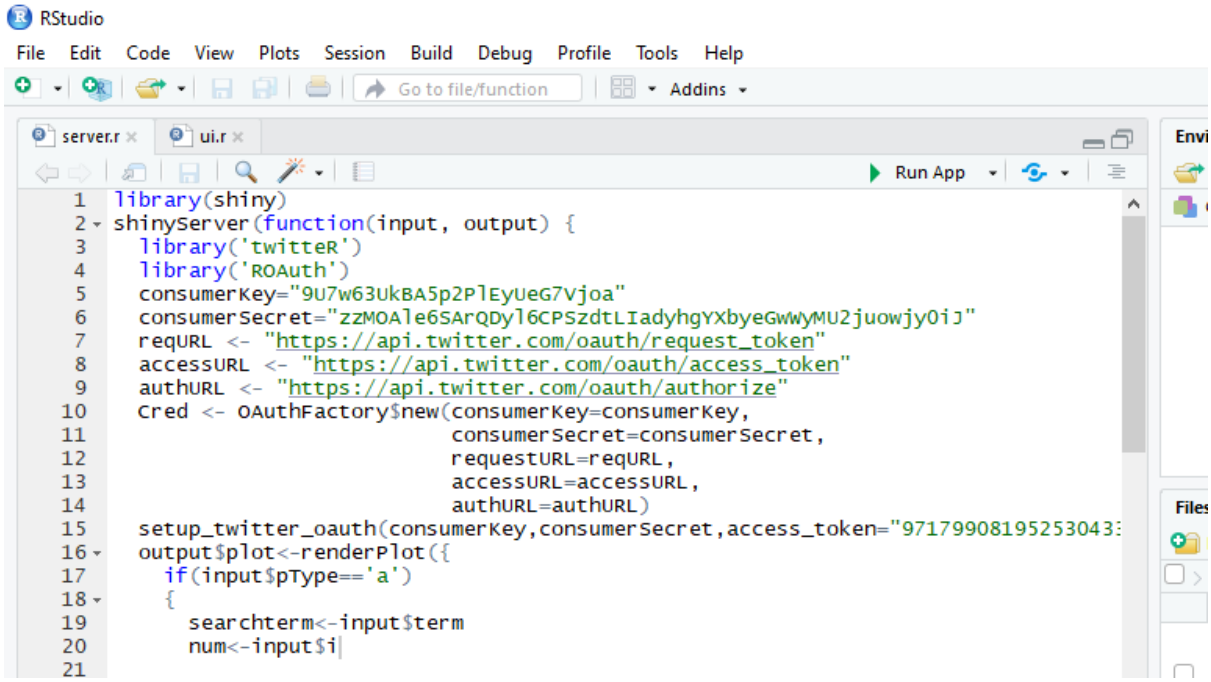
To install in Rstudio

GoTo-->Tools-->click ”Install Packages”-->Type relevant package you need--->click install.

Picture shows installing twitteR package:

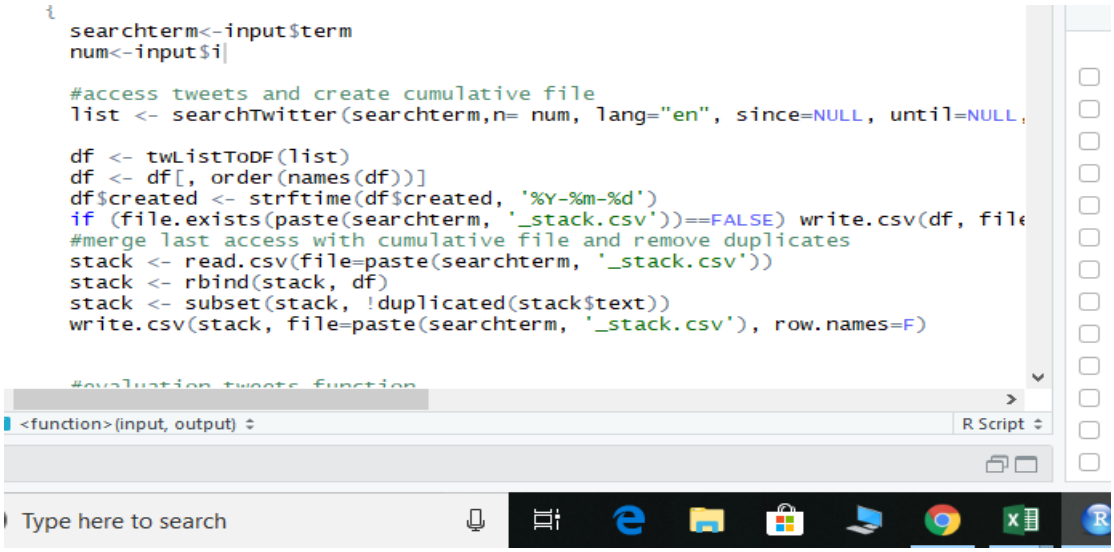


Next we will invoke twitter API using the APP we have created and using the keys and access tokens,we got through that



We will now what format,we got the extract and what all steps do we need to take to clean the data..

2.Cleaning the tweets for further analysis



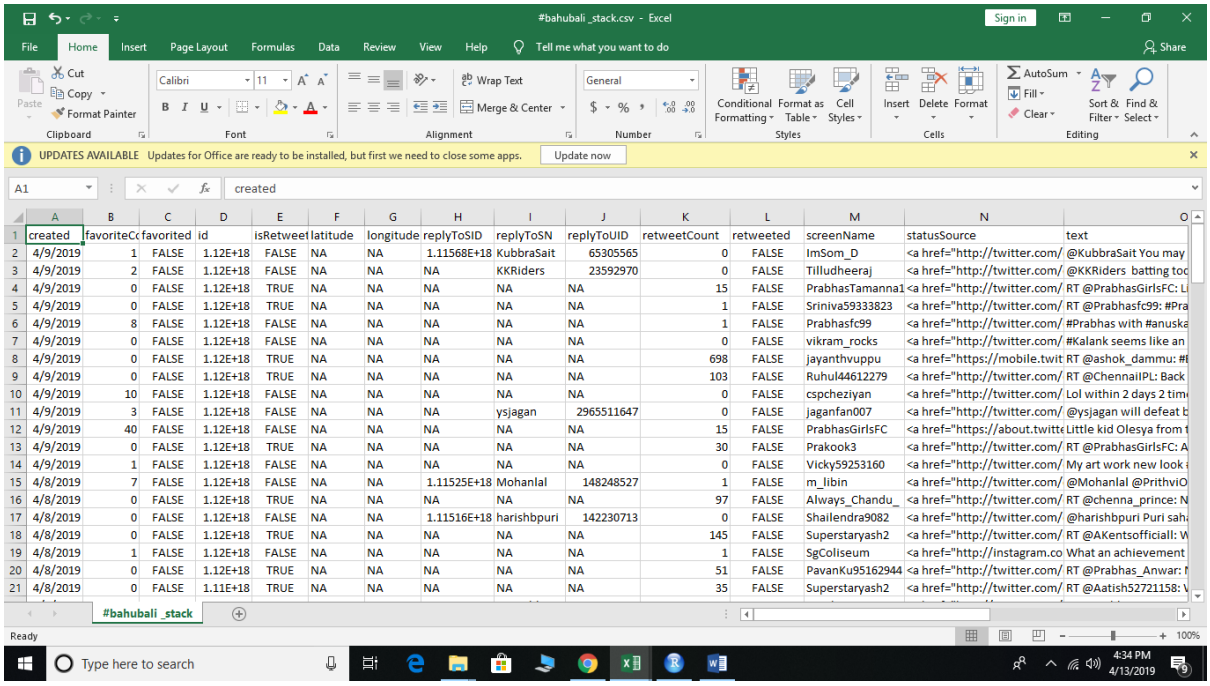
```
searchterm<-input$term
num<-input$i

#access tweets and create cumulative file
list <- searchTwitter(searchterm,n= num, lang="en", since=NULL, until=NULL,

df <- twListToDF(list)
df <- df[, order(names(df))]
df$created <- strftime(df$created, '%Y-%m-%d')
if (file.exists(paste(searchterm, '_stack.csv'))==FALSE) write.csv(df, file
#merge last access with cumulative file and remove duplicates
stack <- read.csv(file=paste(searchterm, '_stack.csv'))
stack <- rbind(stack, df)
stack <- subset(stack, !duplicated(stack$text))
write.csv(stack, file=paste(searchterm, '_stack.csv'), row.names=F)
```

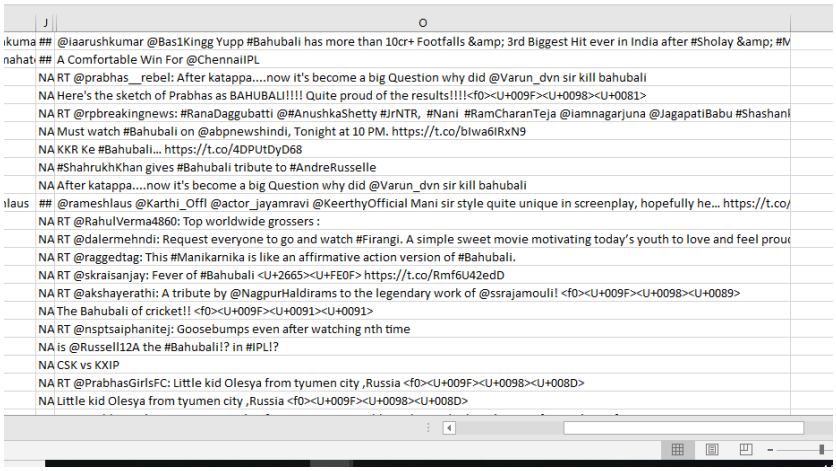

We get a total of using “search Term”…here we used #bahubali..

Snapshot of sample data is shown below…



created	favoriteCc	favorited	id	isRetweet	latitude	longitude	replyToSID	replyToSN	replyToUID	retweetCount	retweeted	screenName	statusSource	text
4/9/2019	1	FALSE	1.12E+18	FALSE	NA	NA	1.11568E+18	KubbraSait	65305565	0	FALSE	ImSom_D	<a href="http://twitter.com/	@KubbraSait You may
4/9/2019	2	FALSE	1.12E+18	FALSE	NA	NA	NA	KKRiders	23592970	0	FALSE	Tilludheeraj	<a href="http://twitter.com/	@KKRiders batting toc
4/9/2019	0	FALSE	1.12E+18	TRUE	NA	NA	NA	NA	NA	15	FALSE	PrabhasTamanna1	<a href="http://twitter.com/	RT @PrabhasGirlsFC: L
4/9/2019	0	FALSE	1.12E+18	TRUE	NA	NA	NA	NA	NA	1	FALSE	Sriniva59333823	<a href="http://twitter.com/	RT @Prabhasfc99: #Pra
4/9/2019	8	FALSE	1.12E+18	FALSE	NA	NA	NA	NA	NA	1	FALSE	Prabhasfc99	<a href="http://twitter.com/	#Prabhas with #anuska
4/9/2019	0	FALSE	1.12E+18	FALSE	NA	NA	NA	NA	NA	0	FALSE	vikram_rocks	<a href="http://twitter.com/	#Kalank seems like an
4/9/2019	0	FALSE	1.12E+18	TRUE	NA	NA	NA	NA	NA	698	FALSE	jayanthvuppu	<a href="https://mobile.twit	RT @ashok_dammu: #I
4/9/2019	0	FALSE	1.12E+18	TRUE	NA	NA	NA	NA	NA	103	FALSE	Ruhul44612279	<a href="http://twitter.com/	RT @ChennaiIPL: Back
4/9/2019	10	FALSE	1.12E+18	FALSE	NA	NA	NA	NA	NA	0	FALSE	cspcheziyan	<a href="http://twitter.com/	Lol within 2 days 2 tim
4/9/2019	3	FALSE	1.12E+18	FALSE	NA	NA	NA	ysjagan	2965511647	0	FALSE	jaganfan007	<a href="http://twitter.com/	@ysjagan will defeat b
4/9/2019	40	FALSE	1.12E+18	FALSE	NA	NA	NA	NA	NA	15	FALSE	PrabhasGirlsFC	<a href="https://about.twitte	Little kid Olesya from t
4/9/2019	0	FALSE	1.12E+18	TRUE	NA	NA	NA	NA	NA	30	FALSE	Prakook3	<a href="http://twitter.com/	RT @PrabhasGirlsFC: A
4/9/2019	1	FALSE	1.12E+18	FALSE	NA	NA	NA	NA	NA	0	FALSE	Vicky59253160	<a href="http://twitter.com/	My art work new look
4/8/2019	7	FALSE	1.12E+18	FALSE	NA	NA	1.11525E+18	Mohanlal	148248527	1	FALSE	m_libin	<a href="http://twitter.com/	@Mohanlal @PrithviO
4/8/2019	0	FALSE	1.12E+18	TRUE	NA	NA	NA	NA	NA	97	FALSE	Always_Chandu_	<a href="http://twitter.com/	RT @chenna_prince: N
4/8/2019	0	FALSE	1.12E+18	FALSE	NA	NA	1.11516E+18	harishbpuri	142230713	0	FALSE	Shailendra9082	<a href="http://twitter.com/	@harishbpuri Puri sah
4/8/2019	0	FALSE	1.12E+18	TRUE	NA	NA	NA	NA	NA	145	FALSE	Superstaryash2	<a href="http://twitter.com/	RT @AKentsofficiall: W
4/8/2019	1	FALSE	1.12E+18	FALSE	NA	NA	NA	NA	NA	1	FALSE	SgColiseum	<a href="http://instagram.co	What an achievement
4/8/2019	0	FALSE	1.12E+18	TRUE	NA	NA	NA	NA	NA	51	FALSE	PavanKu95162944	<a href="http://twitter.com/	RT @Prabhas_Anwar: I
4/8/2019	0	FALSE	1.11E+18	TRUE	NA	NA	NA	NA	NA	35	FALSE	Superstaryash2	<a href="http://twitter.com/	RT @Aatish52721158: V

The field “TEXT” contains the tweet part,hashtags and URL’s,we need to remove hashtags and URL’s from the text field so that we are left only with the main tweet part to run our sentiment analysis..



	J	O
ikuma	##	@iaarushkumar @Bas1Kingg Yupp #Bahubali has more than 10cr+ Footfalls & 3rd Biggest Hit ever in India after #Sholay & #M
nahat	##	A Comfortable Win For @ChennaiIPL
	NA	RT @prabhas__rebel: After katappa....now it's become a big Question why did @Varun_dvn sir kill bahubali
	NA	Here's the sketch of Prabhas as BAHUBALI!!!! Quite proud of the results!!!!<f0><U+009F><U+0098><U+0081>
	NA	RT @rpbreakingnews: #RanaDaggubatti @#AnushkaShetty #JrNTR, #Nani #RamCharanTeja @iamnagarjuna @JagapatiBabu #Shashank
	NA	Must watch #Bahubali on @abpnewshindi, Tonight at 10 PM. https://t.co/bIwa6IRxN9
	NA	KKR Ke #Bahubali... https://t.co/4DPUtDyD68
	NA	#ShahrukhKhan gives #Bahubali tribute to #AndreRusselle
	NA	After katappa....now it's become a big Question why did @Varun_dvn sir kill bahubali
laus	##	@rameshlaus @Karthi_Offl @actor_jayamravi @KeerthyOfficial Mani sir style quite unique in screenplay, hopefully he... https://t.co/
	NA	RT @RahulVerma4860: Top worldwide grossers :
	NA	RT @dalermehndi: Request everyone to go and watch #Firangi. A simple sweet movie motivating today's youth to love and feel proud
	NA	RT @raggedtag: This #Manikarnika is like an affirmative action version of #Bahubali.
	NA	RT @skraisanjay: Fever of #Bahubali <U+2665><U+FE0F> https://t.co/Rmf6U42edD
	NA	RT @akshayerathi: A tribute by @NagpurHaldirams to the legendary work of @ssrajamouli! <f0><U+009F><U+0098><U+0089>
	NA	The Bahubali of cricket!! <f0><U+009F><U+0091><U+0091>
	NA	RT @nsptsaiphanitej: Goosebumps even after watching nth time
	NA	is @Russell12A the #Bahubali!? in #IPL!?
	NA	CSK vs KXIP
	NA	RT @PrabhasGirlsFC: Little kid Olesya from tyumen city ,Russia <f0><U+009F><U+0098><U+008D>
	NA	Little kid Olesya from tyumen city ,Russia <f0><U+009F><U+0098><U+008D>

This contains alot of url’s ,hash Tags and other twitter handles,we remove all these using gsub function.

```
sentence <- iconv(sentence, to='ASCII//TRANSLIT')


sentence <- gsub('[[:punct:]]', "", sentence)
sentence <- gsub('[[:cntrl:]]', "", sentence)
sentence <- gsub('\\d+', "", sentence)
sentence <- tolower(sentence)


word.list <- str_split(sentence, '\\s+')
words <- unlist(word.list)
```

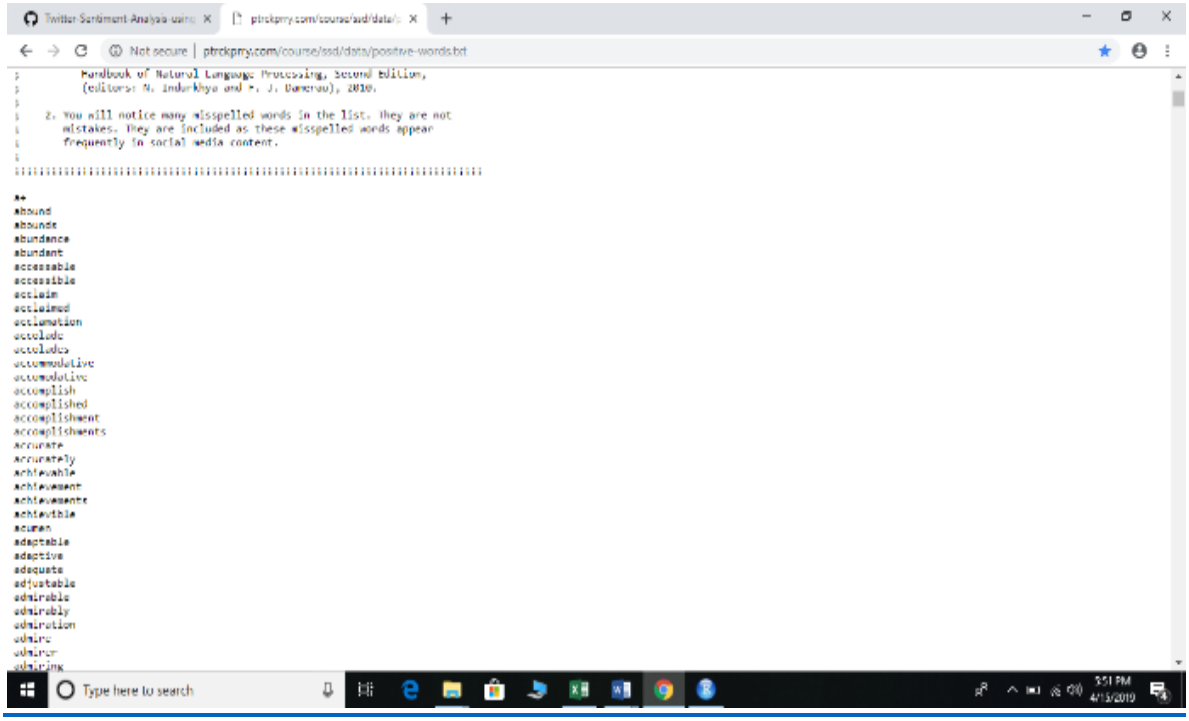
3. Loading Word Database

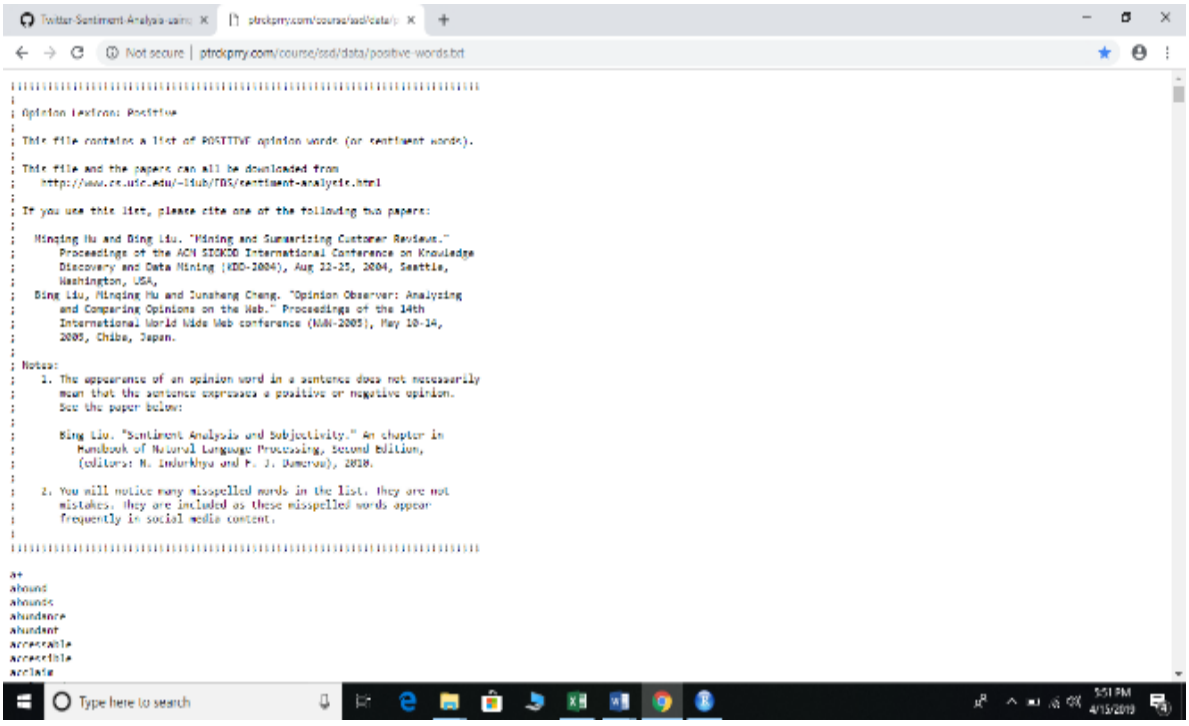
A database, created by Hui Lui containing positive and negative words, is loaded into R. This is used for Lexical Analysis, where the words in the tweets are compared with the words in the database and the sentiment is predicted.

For movie tweets, Naive Bayes Machine Learning Algorithm is used. AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated. The version used is: AFINN-111: Newest version with 2477 words and phrases.

Positive dictionary:

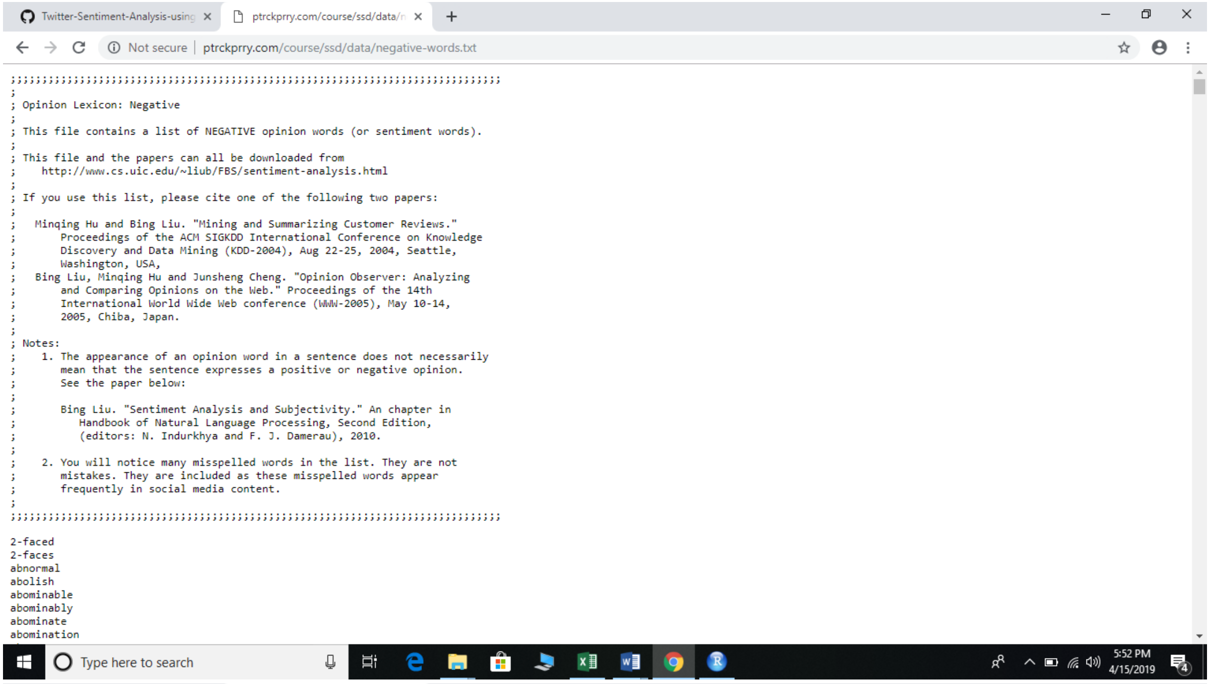
http://ptrckprry.com/course/ssd/data/positive-words.txt





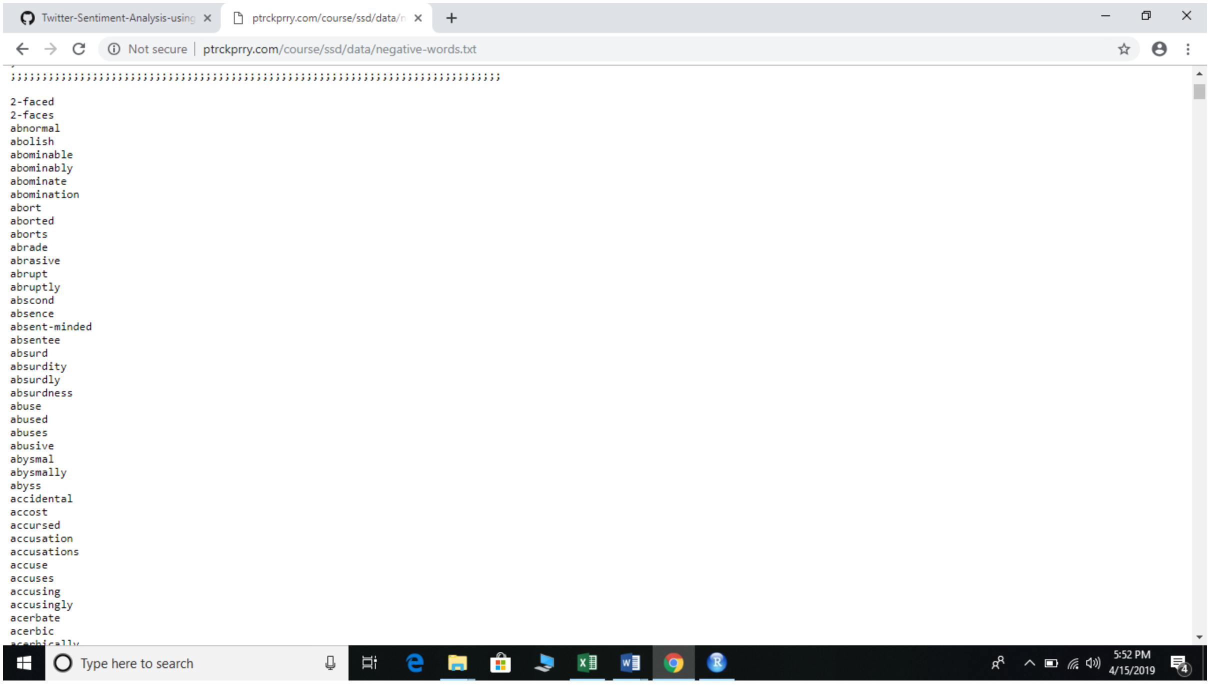
Negative dictionary:

http://ptrckprry.com/course/ssd/data/negative-words.txt



```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;
; Opinion Lexicon: Negative
;
; This file contains a list of NEGATIVE opinion words (or sentiment words).
;
; This file and the papers can all be downloaded from
;    http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
;
; If you use this list, please cite one of the following two papers:
;
;   Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."
;       Proceedings of the ACM SIGKDD International Conference on Knowledge
;       Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle,
;       Washington, USA,
;   Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing
;       and Comparing Opinions on the Web." Proceedings of the 14th
;       International World Wide Web conference (WWW-2005), May 10-14,
;       2005, Chiba, Japan.
;
; Notes:
;    1. The appearance of an opinion word in a sentence does not necessarily
;       mean that the sentence expresses a positive or negative opinion.
;       See the paper below:
;
;       Bing Liu. "Sentiment Analysis and Subjectivity." An chapter in
;          Handbook of Natural Language Processing, Second Edition,
;          (editors: N. Indurkhya and F. J. Damerau), 2010.
;
;    2. You will notice many misspelled words in the list. They are not
;       mistakes. They are included as these misspelled words appear
;       frequently in social media content.
;
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

2-faced
2-faces
abnormal
abolish
abominable
abominably
abominate
abomination
```



```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

2-faced
2-faces
abnormal
abolish
abominable
abominably
abominate
abomination
abort
aborted
aborts
abrade
abrasive
abrupt
abruptly
abscond
absence
absent-minded
absentee
absurd
absurdity
absurdly
absurdness
abuse
abused
abuses
abusive
abysmal
abysmally
abyss
accidental
accost
accursed
accusation
accusations
accuse
accuses
accusing
accusingly
acerbate
acerbic
```

4. Algorithm used

● Lexical Analysis: By comparing uni-grams to the pre-loaded word database, the tweet is assigned sentiment score - positive, negative or neutral and overall score is calculated.

5. Calculating percentage

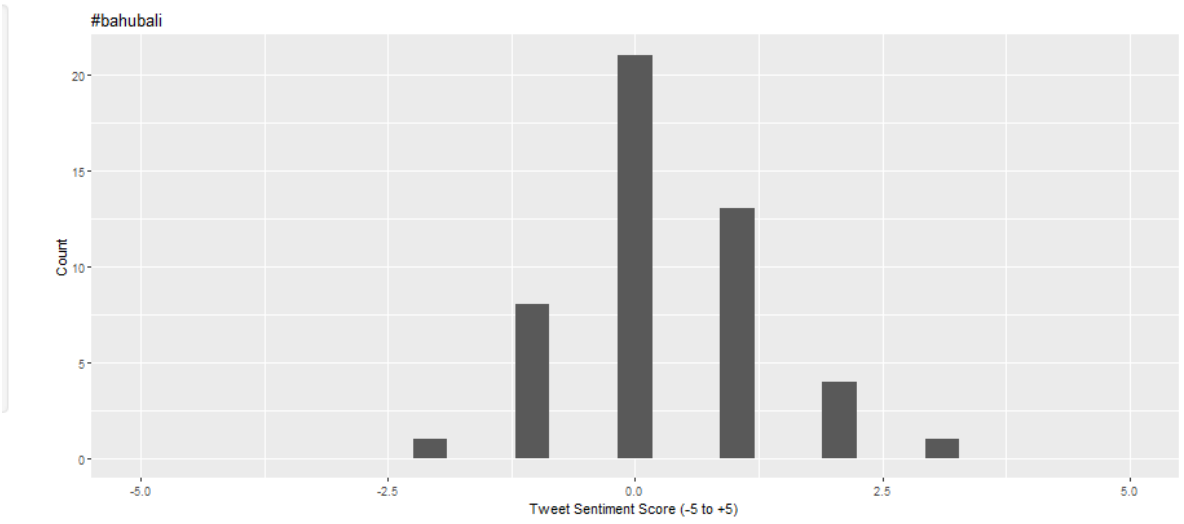
I have presented the scores, the tweets as well as the percentage of positive/negative emotion in the text. This calculated using simple arithmetic to understand the overall sentiment in a more bettermanner.

Evaluating score:

```
pos.matches =match(words, pos.words)
neg.matches = match(words, neg.words)
pos.matches = !is.na(pos.matches)
neg.matches = !is.na(neg.matches)
    score =sum(pos.matches) - sum(neg.matches)
    return(score)
```

6. Histogram tab : histogram plot

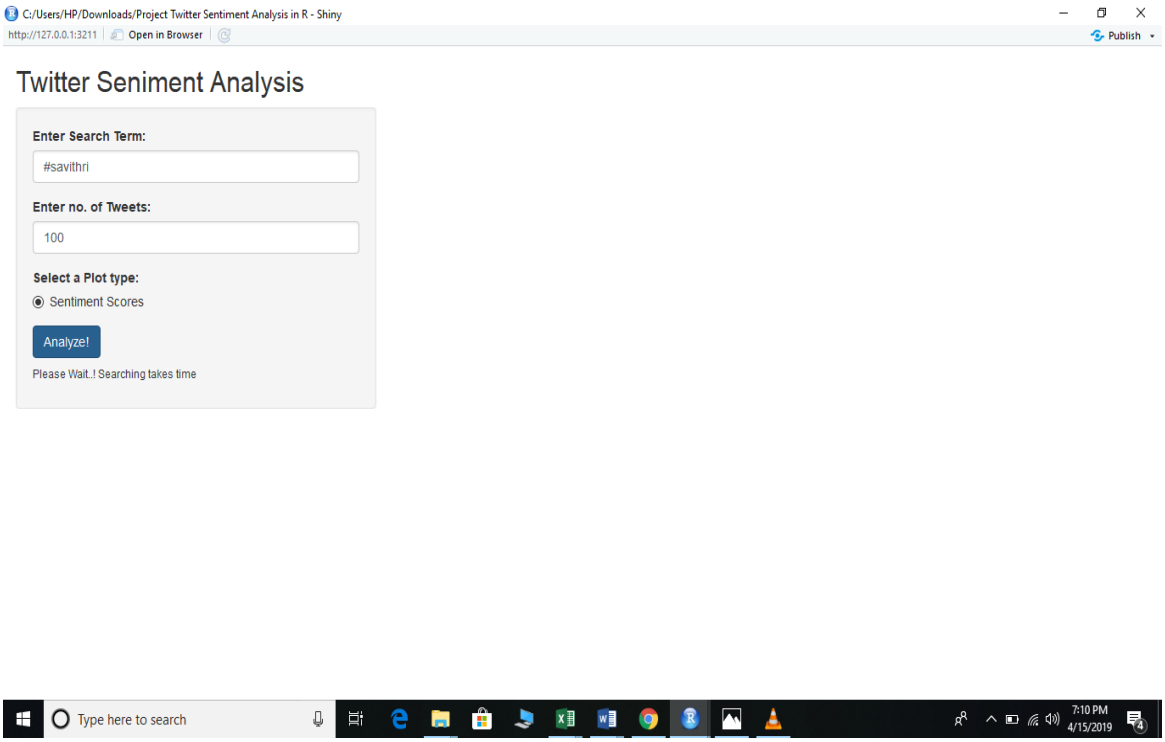
Histograms of positive, negative and overall score are found under the Histogram tab for graphically analyzing the intensity of emotion in the tweeters.



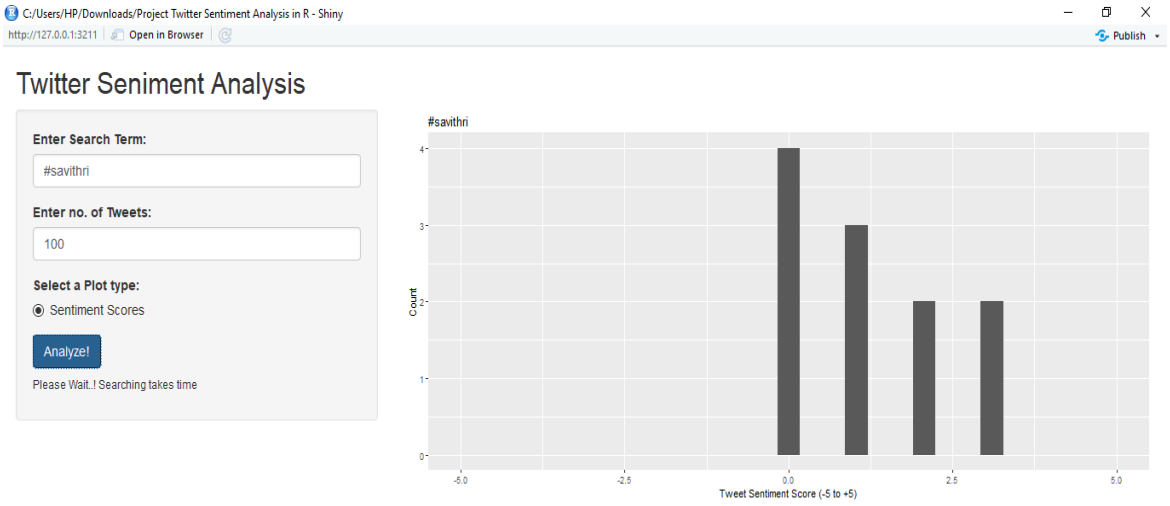
SCREENSHOTS:

WEB PAGE OF OUR PROJECT:

Sentiment analysis on #savithri…



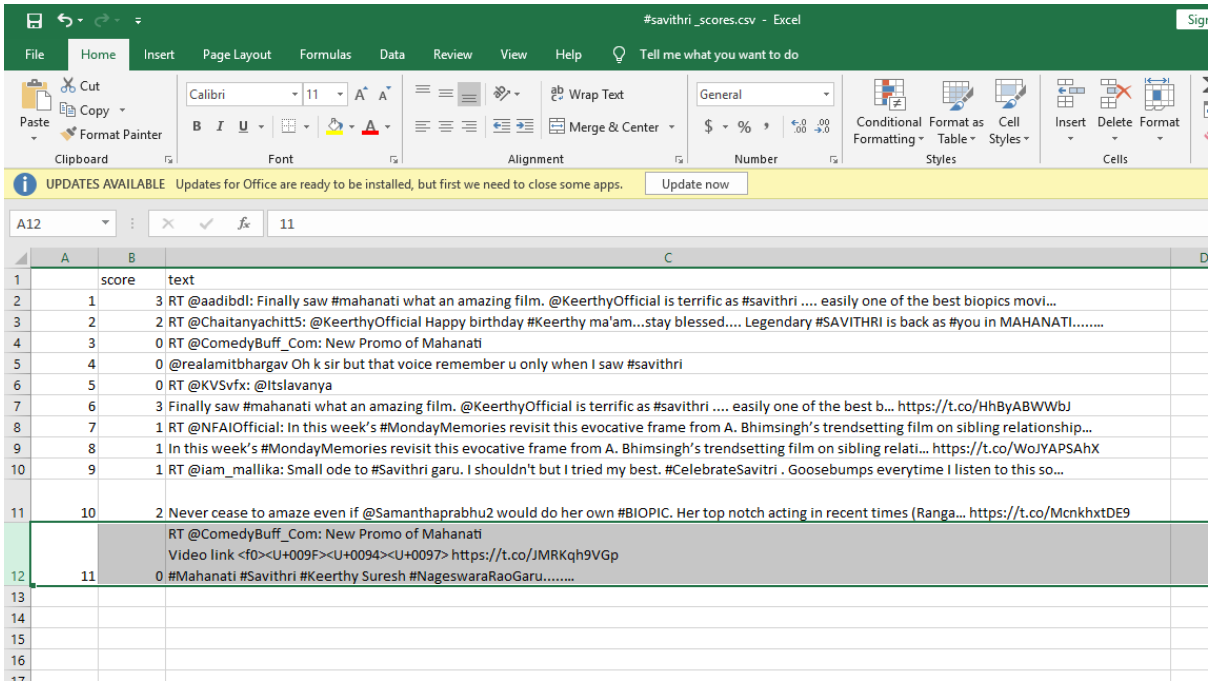
Output:



Files Generated: After analysis 2 excel files are generated.

Scores.csv file:

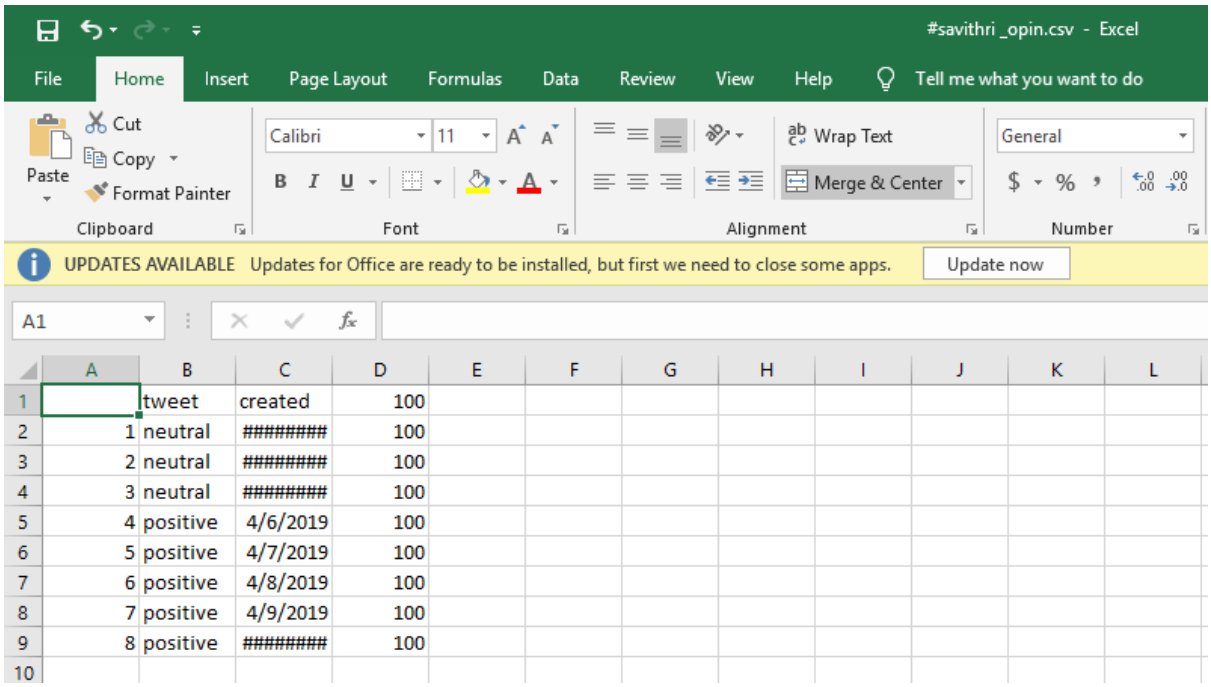
In this sheet,each and every tweets along with their scores are displayed



	A	B	C
1		score	text
2	1	3	RT @aadibdl: Finally saw #mahanati what an amazing film. @KeerthyOfficial is terrific as #savithri .... easily one of the best biopics movi...
3	2	2	RT @Chaitanyachitt5: @KeerthyOfficial Happy birthday #Keerthy ma'am...stay blessed.... Legendary #SAVITHRI is back as #you in MAHANATI.....
4	3	0	RT @ComedyBuff_Com: New Promo of Mahanati
5	4	0	@realamitbhargav Oh k sir but that voice remember u only when I saw #savithri
6	5	0	RT @KVSvfx: @Itslavanya
7	6	3	Finally saw #mahanati what an amazing film. @KeerthyOfficial is terrific as #savithri .... easily one of the best b... https://t.co/HhByABWWbJ
8	7	1	RT @NFAIOfficial: In this week's #MondayMemories revisit this evocative frame from A. Bhimsingh's trendsetting film on sibling relationship...
9	8	1	In this week's #MondayMemories revisit this evocative frame from A. Bhimsingh's trendsetting film on sibling relati... https://t.co/WoJYAPSAhX
10	9	1	RT @iam_mallika: Small ode to #Savithri garu. I shouldn't but I tried my best. #CelebrateSavitri . Goosebumps everytime I listen to this so...
11	10	2	Never cease to amaze even if @Samanthaprabhu2 would do her own #BIOPIC. Her top notch acting in recent times (Ranga... https://t.co/McnkhxtDE9
12	11	0	RT @ComedyBuff_Com: New Promo of Mahanati Video link <f0><U+009F><U+0094><U+0097> https://t.co/JMRKqh9VGp #Mahanati #Savithri #Keerthy Suresh #NageswaraRaoGaru.....
13			
14			
15			
16			

Opinion.csv file:

In this sheet,polarity of each and every scored tweets are displayed.

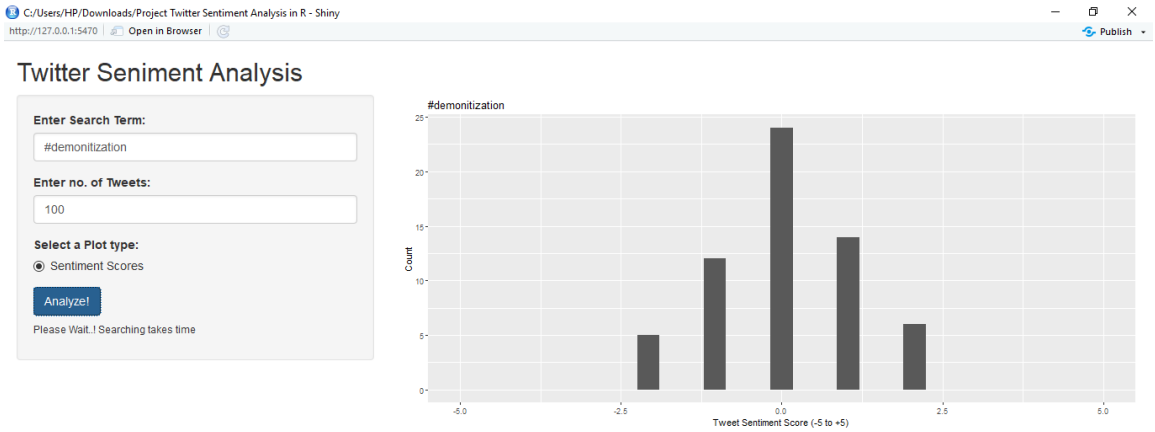


	A	B	C	D	E	F	G	H	I	J	K	L
1		tweet	created	100								
2	1	neutral	########	100								
3	2	neutral	########	100								
4	3	neutral	########	100								
5	4	positive	4/6/2019	100								
6	5	positive	4/7/2019	100								
7	6	positive	4/8/2019	100								
8	7	positive	4/9/2019	100								
9	8	positive	########	100								
10												

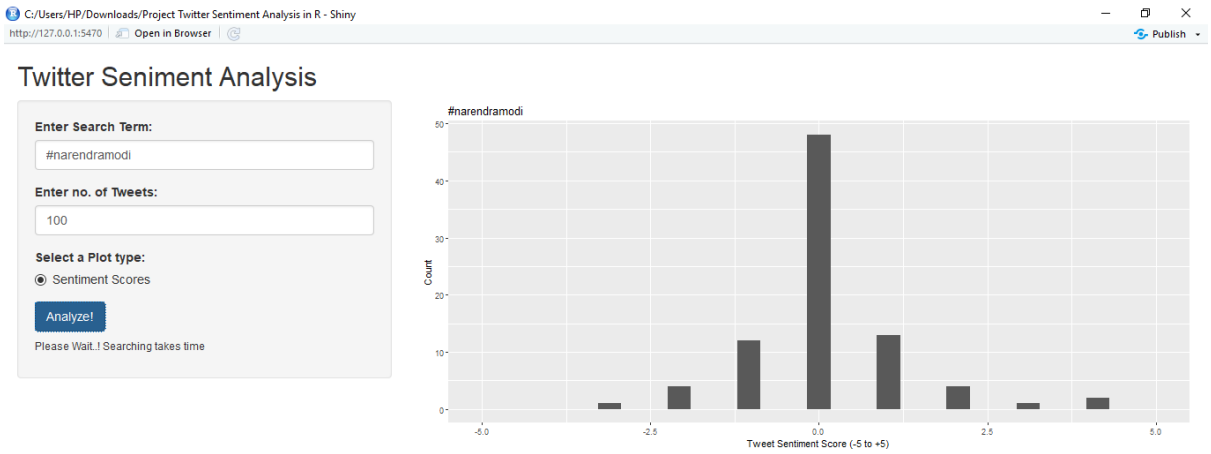
SCREENSHOTS:

Some of our works:

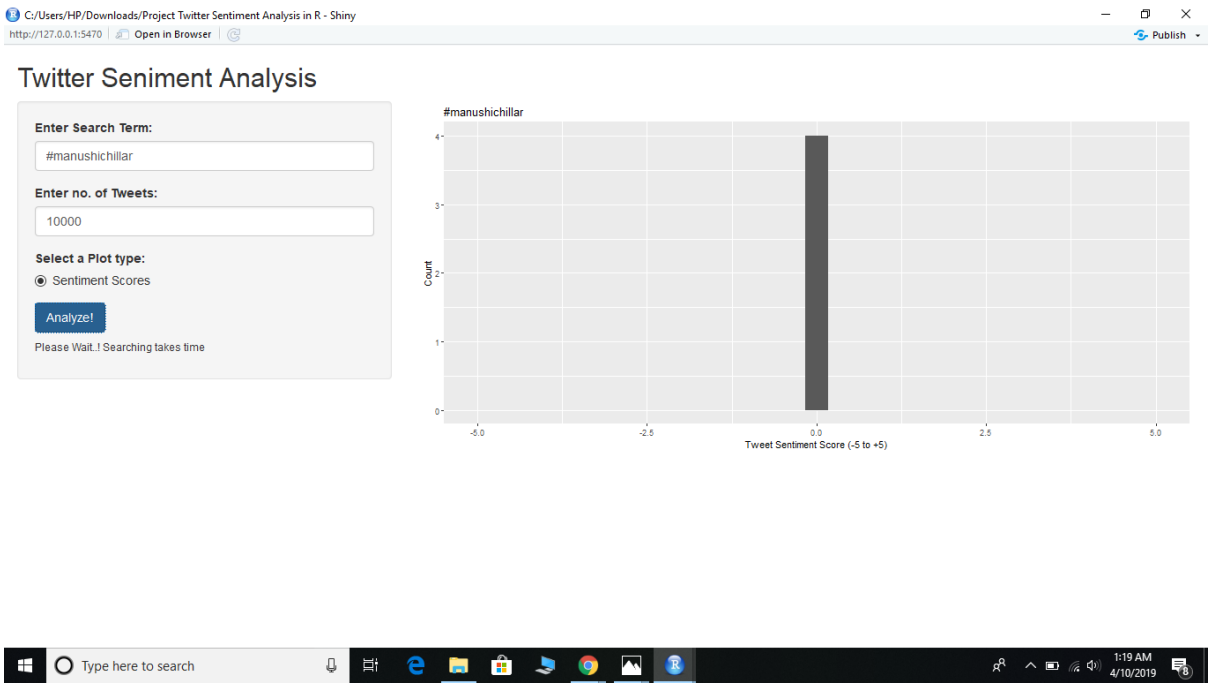
#demonitization



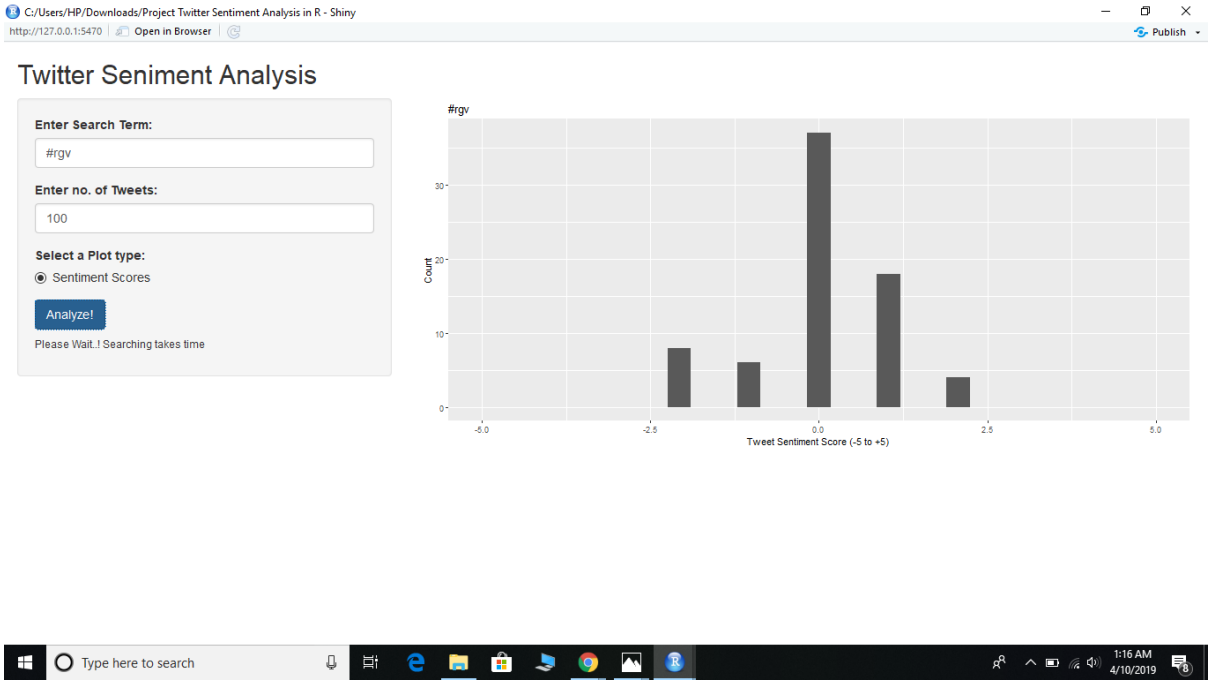
#narendramodi



#manushichillar



#rgv



APPLICATIONS OF SENTIMENT ANALYSIS

Sentiment Analysis has many applications in various Fields.

1.Applications that use Reviewsfrom Websites:

Today Internet has a large collection of reviews and feedbacks on almost everything. This includes product reviews, feedbacks on political issues, comments about services, etc. Thus there is a need for a sentiment analysis system that can extract sentiments about a particular product or services. It will help us to automate in provision of feedback or rating for the given product, item, etc. This would serve the needs of both the users and the vendors.

2. Applications as a Sub-component Technology

A sentiment predictor system can be helpful in recommender systems as well. The recommender system will not recommend items that receive a lot of negative feedback or fewer ratings.

In online communication, we come across abusive language and other negative elements. These can be detected simply by identifying a highly negative sentiment and correspondingly taking action against it. 3. *Applications in Business Intelligence*

It has been observed that people nowadays tend to look upon reviews of products which are available online before they buy them. And for many businesses, the online opinion decides the success or failure of their product. Thus, Sentiment Analysis plays an important role in businesses. Businesses also wish to extract sentiment from the online reviews in order to improve their products and in turn their reputation and help in customer satisfaction .

4. ***Applications across Domains:***

Recent researches in sociology and other fields like medical, sports have also been benefitted by Sentiment Analysis that show trends in human emotions especially on social media.

5. ***Applications In Smart Homes***

Smart homes are supposed to be the technology of the future. In future entire homes would be networked and people would be able to control any part of the home using a tablet device. Recently there has been lot of research going on Internet of Things (IoT). Sentiment Analysis would also find its way in IoT. Like for example, based on the current sentiment or emotion of the user, the home could alter its ambiance to create a soothing and peaceful environment.

Sentiment Analysis can also be used in trend prediction. By tracking public views, important data regarding sales trends and customer satisfaction can be extracted.

FUTURE WORK

- Detect sarcasm in tweets
- Analyse images for emotions.
- Add Telugu words to dataset.
- Find no of mentions of n particular organizations.
- Parallelizing code.
- Apply better Machine Learning Algorithms (Like Support Vector Machine).

REFERENCES

- https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
- https://www.quora.com/How-can-I-read-Twitter-data-directly-in-R
- https://www.r-bloggers.com/emoticons-decoder-for-social-media-sentimentanalysis-in-r/
- https://eight2late.wordpress.com/2015/11/06/a-gentle-introduction-to-naive-bayes-classification-using-r/