A PRODUCT DEVELOPMENT PROJECT REPORT ON

TWITTER SENTIMENT ANALYSIS & VISUALIZATION

Submitted by:

Sonali Agrawal
114CS0634

Asmita Poddar
114CS0375

National Institute of Technology, Rourkela

Project Guide:

Dr. Ashok Kumar Turuk

Start Date of project: 20th August 2016

End date of project: 26th October, 2016

CONTENTS

PREFACE

This report documents the work done during product development lab in National Institute of Technology, Rourkela under the supervision of Dr. Ashok Kumar Turuk.

The report first shall give an overview of Twitter Sentiment Analysis and visualization and then about each part in detail. Report shall also elaborate on the future works which can be persuaded as an advancement of the current work.

Sonali Agrawal	**Asmita Poddar**
114CS0634	**114CS0375**

ACKNOWLEDGEMENT

This work was done as a part of the Product Development Laboratory course undertaken by the authors at National Institute of Technology, Rourkela during the period dated 20th August, 2016 to 26th October, 2016.

The authors would like to acknowledge the constant guidance, encouragement and supervision rendered by Dr. Ashok Kumar Turuk.

The authors also express their sincere gratitude to National Institute of Technology and access the library and computational facilities on campus.

We perceive this opportunity as a big milestone in career development. We will strive to use gained skills and knowledge in the best possible way, and will continue to work on their improvement, in order to attain desired career objectives.

INTRODUCTION

CONTEXT

- Numerous outlets available for individuals to express opinions and emotions...positive, negative, and neutral.
- Need to promote positive news, react to the negative, and move the needle favorably on neutral news....as near real-time as possible
- Mining high volume, high velocity data for meaningful insights is not easy!...too much, too fast
- Similar challenges exist across all industries/verticals

WHY ANALYTICS?

◦ What is trending positively/negatively over a period of time and why?

◦ Who is being talked about, where, and why?

◦ What college is being talked about?

◦ What topics are being discussed the most?

◦ Who is being talked about most positively?

◦ What are the best sources for positive exposure?

◦ What is the geographic location of the comments?

WHY 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, barplots, timelines, word clouds, histograms and pie charts can be used for visualization.

End Result: Informed Strategies, Improved Performance

WHAT IS 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.

Twitter Sentiment Analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the 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 – Semantria 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 ahead 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.

OVERVIEW

Tweets are imported using R and the data is cleaned by removing emoticons and URLs. Lexical Analysis as well as Naive Bayes Classifier is used to predict the sentiment of tweets and subsequently express the opinion graphically through ggplots, histogram, pie chart, wordcloud and tables. The front end has been created using the Shiny App.

SYSTEM REQUIREMENTS

- Installation of R
- Twitter Authentication to access API

FEATURES

1. **Extraction of Tweets**

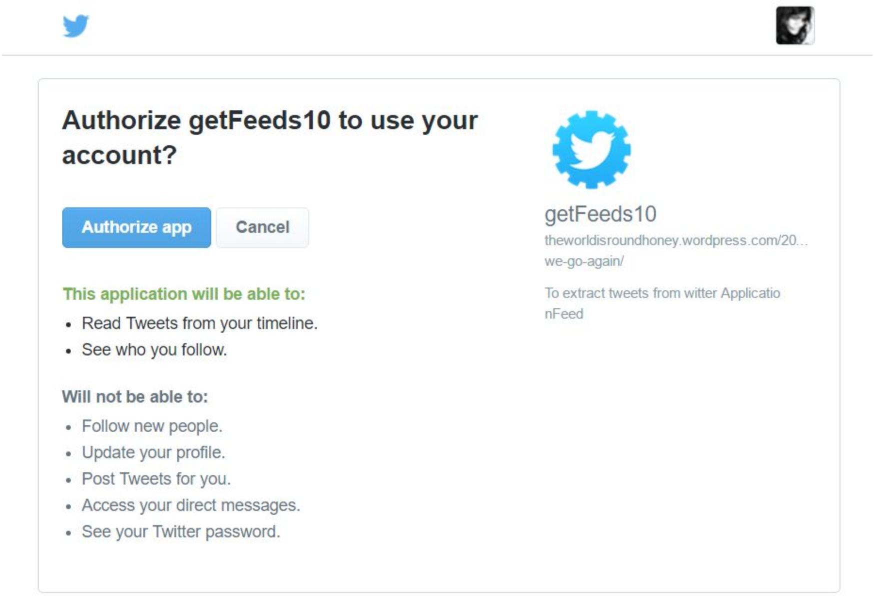
   (i) Create twitter application

   (ii) twitteR - Provides an interface to the Twitter web API

   (iii) ROAuth - R Interface For OAuth

   (iv)Create twitter authenticated credential object(using key from step (ii) and cacert.pem certificate): It is done using consumer key, consumer secret, access token, access secret.

   (v) During authentication, we are redirected to a URL automatically where we click on Authorize app as shown in the image below and enter the unique 7-digit number to get linked to the account from which feeds are being taken.



Authorize getFeeds10 to use your account?

This application will be able to:

- Read Tweets from your timeline.
- See who you follow.

Will not be able to:

- Follow new people.
- Update your profile.
- Post Tweets for you.
- Access your direct messages.
- See your Twitter password.

getFeeds10
theworldisroundhoney.wordpress.com/20...
we-go-again/

To extract tweets from witter Applicatio
nFeed

2. Cleaning Tweets

The tweets are cleaned in R by removing:

- Extra punctuation
- Stop words (Most commonly used words in a language like *the*, *is*, *at*, *which*, and *on*.)
- Redundant Blank spaces
- Emoticons
- URLS

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.

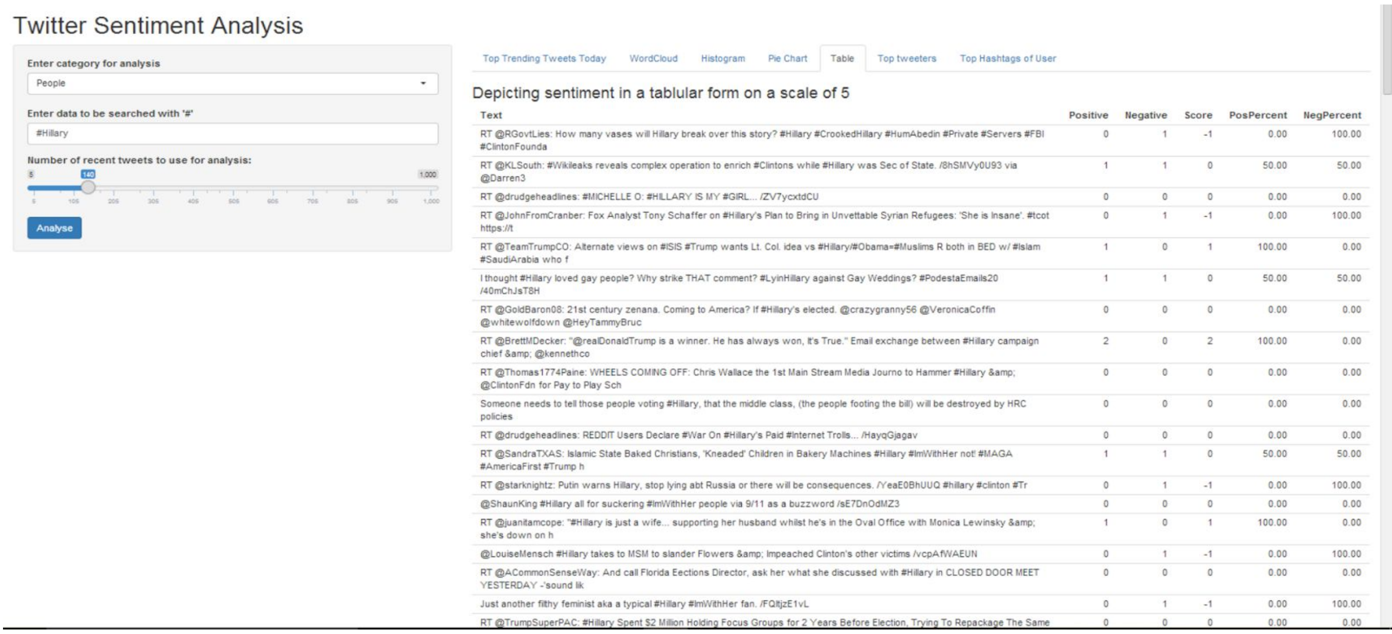
For movie tweets, Naive Bayes Machine Learning Algorithm is used. AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated. The version used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

4. Algorithms 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.
- **Naive Bayes Machine Learning Algorithm:** Training data sets are used to teach the machine what kind of sentences are categorized as positive and what kind are categorized as negative. On arrival of a new tweet or sentence, the machine uses this algorithm to give the correct category to the new data and adds level to the emotion.

5. Calculating percentage

In the table tab of our Shiny Web app as shown below, we have presented the scores, the tweets as well as the percentage of positive/negative emotion in the text. Th is calculated using simple arithmetic to understand the overall sentiment in a more better manner.



Twitter Sentiment Analysis

Enter category for analysis: People

Enter data to be searched with '#': #Hillary

Number of recent tweets to use for analysis: 140

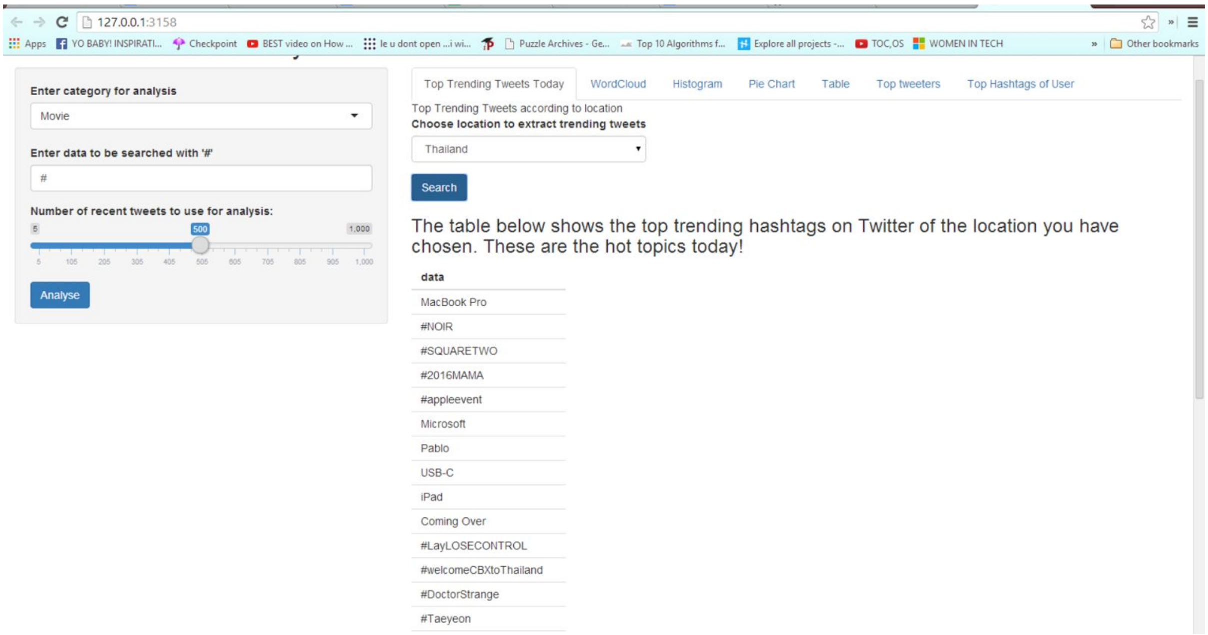
Depicting sentiment in a tablular form on a scale of 5

Text	Positive	Negative	Score	PosPercent	NegPercent
RT @RGovtLies: How many vases will Hillary break over this story? #Hillary #CrookedHillary #HumAbedin #Private #Servers #FBI #ClintonFounda	0	1	-1	0.00	100.00
RT @KLSouth: #Wikileaks reveals complex operation to enrich #Clintons while #Hillary was Sec of State. /8hSMVy0U93 via @Darren3	1	1	0	50.00	50.00
RT @drudgeheadlines: #MICHELLE O: #HILLARY IS MY #GIRL... /ZV7ycxtdCU	0	0	0	0.00	0.00
RT @JohnFromCranber: Fox Analyst Tony Schaffer on #Hillary's Plan to Bring in Unvettable Syrian Refugees: 'She is Insane'. #tcot https://t	0	1	-1	0.00	100.00
RT @TeamTrumpCO: Alternate views on #ISIS #Trump wants Lt. Col. idea vs #Hillary/#Obama=#Muslims R both in BED w/ #Islam #SaudiArabia who f	1	0	1	100.00	0.00
I thought #Hillary loved gay people? Why strike THAT comment? #LyinHillary against Gay Weddings? #PodestaEmails20 /40mChJsT8H	1	1	0	50.00	50.00
RT @GoldBaron08: 21st century zenana. Coming to America? If #Hillary's elected. @crazygranny56 @VeronicaCoffin @whitewolfdown @HeyTammyBruc	0	0	0	0.00	0.00
RT @BrettMDecker: "@realDonaldTrump is a winner. He has always won, It's True." Email exchange between #Hillary campaign chief &amp; @kennethco	2	0	2	100.00	0.00
RT @Thomas1774Paine: WHEELS COMING OFF: Chris Wallace the 1st Main Stream Media Journo to Hammer #Hillary &amp; @ClintonFdn for Pay to Play Sch	0	0	0	0.00	0.00
Someone needs to tell those people voting #Hillary, that the middle class, (the people footing the bill) will be destroyed by HRC policies	0	0	0	0.00	0.00
RT @drudgeheadlines: REDDIT Users Declare #War On #Hillary's Paid #Internet Trolls... /HayqGjagav	0	0	0	0.00	0.00
RT @SandraTXAS: Islamic State Baked Christians, 'Kneaded' Children in Bakery Machines #Hillary #ImWithHer not! #MAGA #AmericaFirst #Trump h	1	1	0	50.00	50.00
RT @starknightz: Putin warns Hillary, stop lying abt Russia or there will be consequences. /YeaE0BhUUQ #hillary #clinton #Tr	0	1	-1	0.00	100.00
@ShaunKing #Hillary all for suckering #ImWithHer people via 9/11 as a buzzword /sE7DnOdMZ3	0	0	0	0.00	0.00
RT @juanitamcope: "#Hillary is just a wife... supporting her husband whilst he's in the Oval Office with Monica Lewinsky &amp; she's down on h	1	0	1	100.00	0.00
@LouiseMensch #Hillary takes to MSM to slander Flowers &amp; Impeached Clinton's other victims /vcpAfWAEUN	0	1	-1	0.00	100.00
RT @ACommonSenseWay: And call Florida Eections Director, ask her what she discussed with #Hillary in CLOSED DOOR MEET YESTERDAY -'sound lik	0	0	0	0.00	0.00
Just another filthy feminist aka a typical #Hillary #ImWithHer fan. /FQltjzE1vL	0	1	-1	0.00	100.00
RT @TrumpSuperPAC: #Hillary Spent $2 Million Holding Focus Groups for 2 Years Before Election, Trying To Repackage The Same	0	0	0	0.00	0.00

6. Top Trending Tweets Today tab: Table

The table is shown which displays the top trending hashtags on Twitter of the location that has been selected.

A **WOEID (Where On Earth IDentifier)** is a unique 32-bit reference identifier, which is generated, and R uses the WOEID of the selected place to obtain the trending hashtags from that location.



Enter category for analysis: Movie

Enter data to be searched with '#': #

Number of recent tweets to use for analysis: 500

Analyse

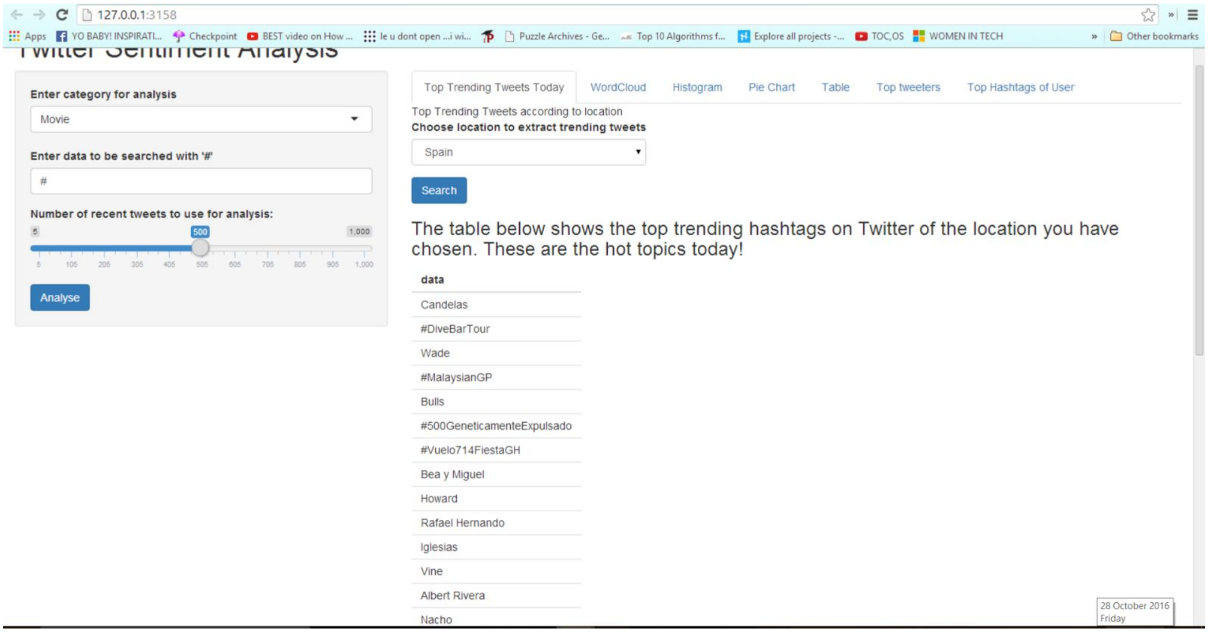
Top Trending Tweets according to location

Choose location to extract trending tweets: Thailand

Search

The table below shows the top trending hashtags on Twitter of the location you have chosen. These are the hot topics today!

data
MacBook Pro
#NOIR
#SQUARETWO
#2016MAMA
#appleevent
Microsoft
Pablo
USB-C
iPad
Coming Over
#LayLOSECONTROL
#welcomeCBXtoThailand
#DoctorStrange
#Taeyeon



Enter category for analysis: Movie

Enter data to be searched with '#': #

Number of recent tweets to use for analysis: 500

Analyse

Top Trending Tweets according to location

Choose location to extract trending tweets: Spain

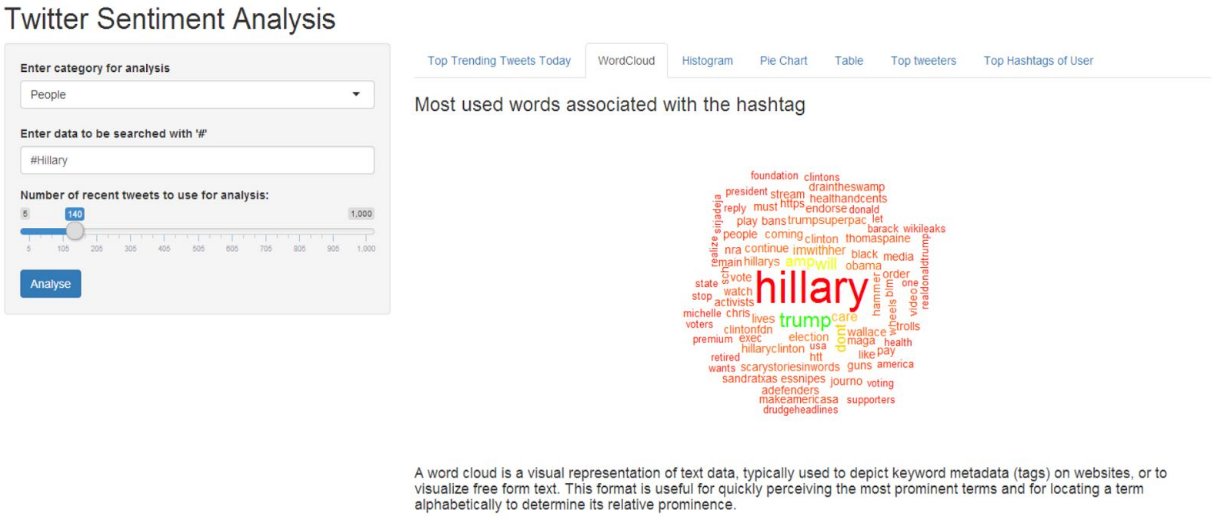
Search

The table below shows the top trending hashtags on Twitter of the location you have chosen. These are the hot topics today!

data
Candelas
#DiveBarTour
Wade
#MalaysianGP
Bulls
#500GeneticamenteExpulsado
#Vuelo714FiestaGH
Bea y Miguel
Howard
Rafael Hernando
Iglesias
Vine
Albert Rivera
Nacho

7 . Word Cloud tab : wordcloud

A word cloud is a visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence. We have used tm and wordcloud package to depict the most used words associated with the hashtag in a pictorial representation under the Wordcloud tab.



Twitter Sentiment Analysis

Enter category for analysis: People

Enter data to be searched with '#': #Hillary

Number of recent tweets to use for analysis: 140

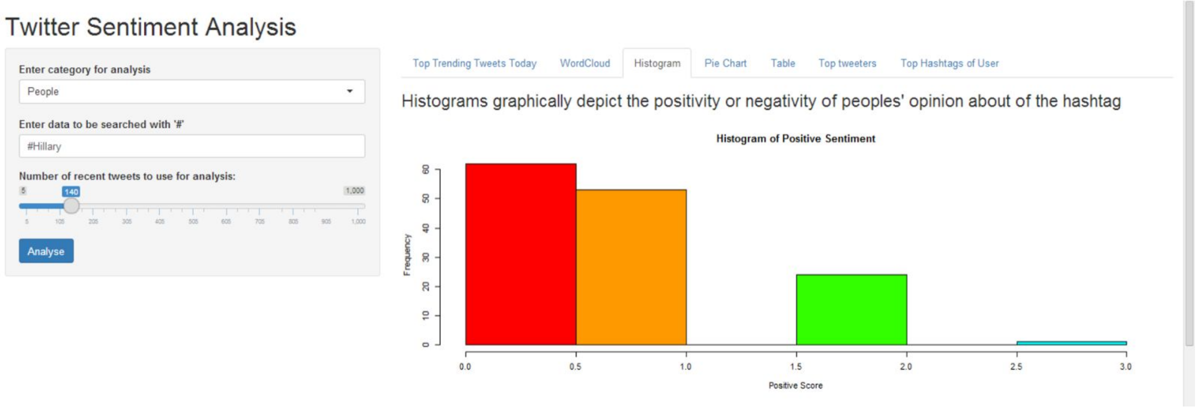
Most used words associated with the hashtag

A word cloud is a visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence.

Top Trending Tweets Today
WordCloud
Histogram
Pie Chart
Table
Top tweeters
Top Hashtags of User
5
1,000
5 105 205 305 405 505 605 705 805 905 1,000
Analyse
foundation clintons
draintheswamp
president stream healthandcents
sirjadeja reply must https endorse donald
play bans trumpsuperpac let
barack wikileaks
people coming clinton thomaspaine
realize nra continue imwithher black media realdonaldtrump
main hillarys amp will obama
sch vote order
state hillary hammer blm one
stop watch video
activists
michelle chris lives trump care wheels
voters clintonfdn dont wallace trolls
premium exec election maga health
hillaryclinton usa
retired htt like pay
wants scarystoriesinwords guns america
sandratxas essnipes journo voting
adefenders
makeamericasa supporters
drudgeheadlines

8. 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.



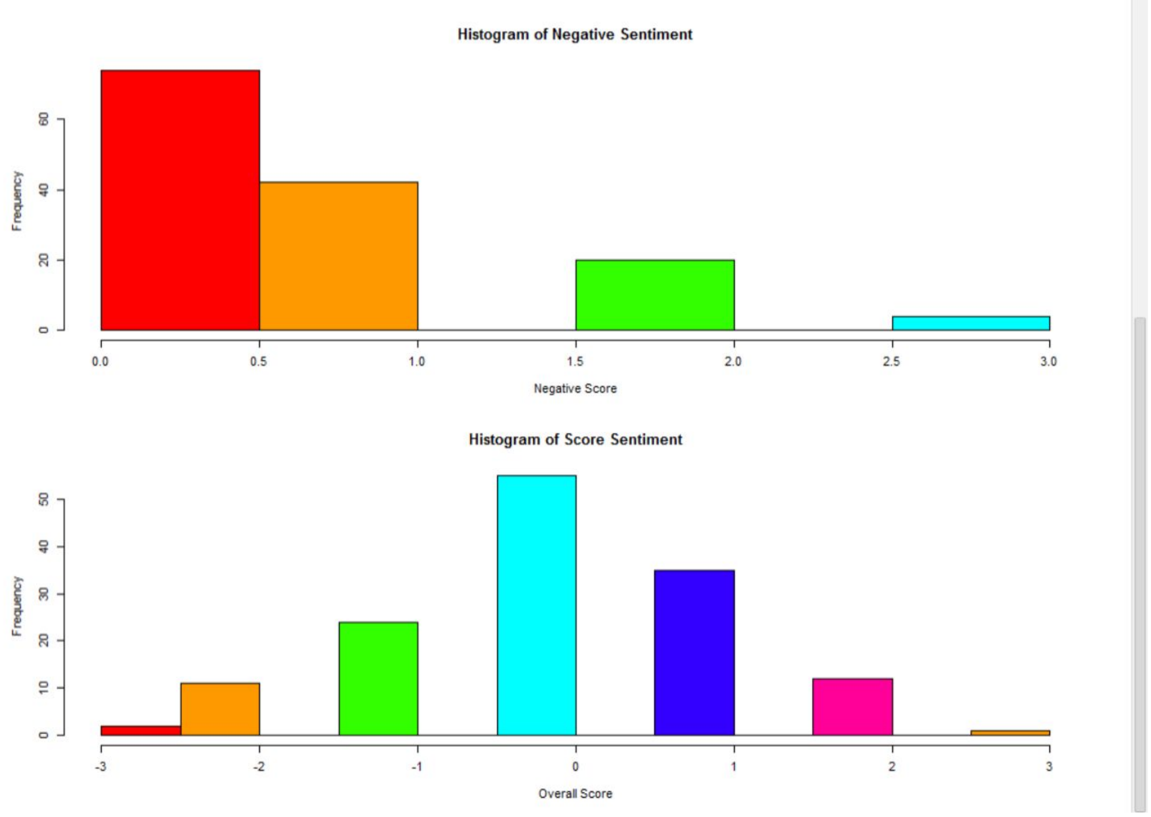
Twitter Sentiment Analysis

Enter category for analysis: People

Enter data to be searched with '#': #Hillary

Number of recent tweets to use for analysis: 140

Histograms graphically depict the positivity or negativity of peoples' opinion about of the hashtag

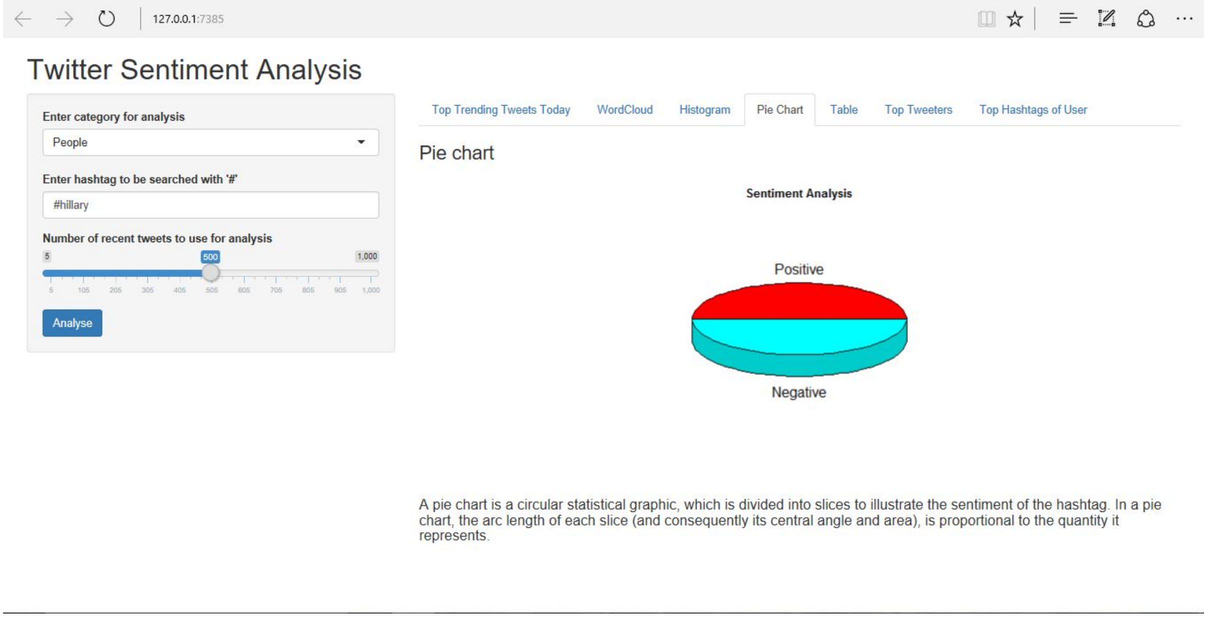


Histogram of Negative Sentiment

Histogram of Score Sentiment

9. Pie Chart tab : pie chart plot

A pie chart is a circular statistical graphic, which is divided into slices to illustrate the sentiment of the hashtag. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.



Twitter Sentiment Analysis

Enter category for analysis: People

Enter hashtag to be searched with '#': #hillary

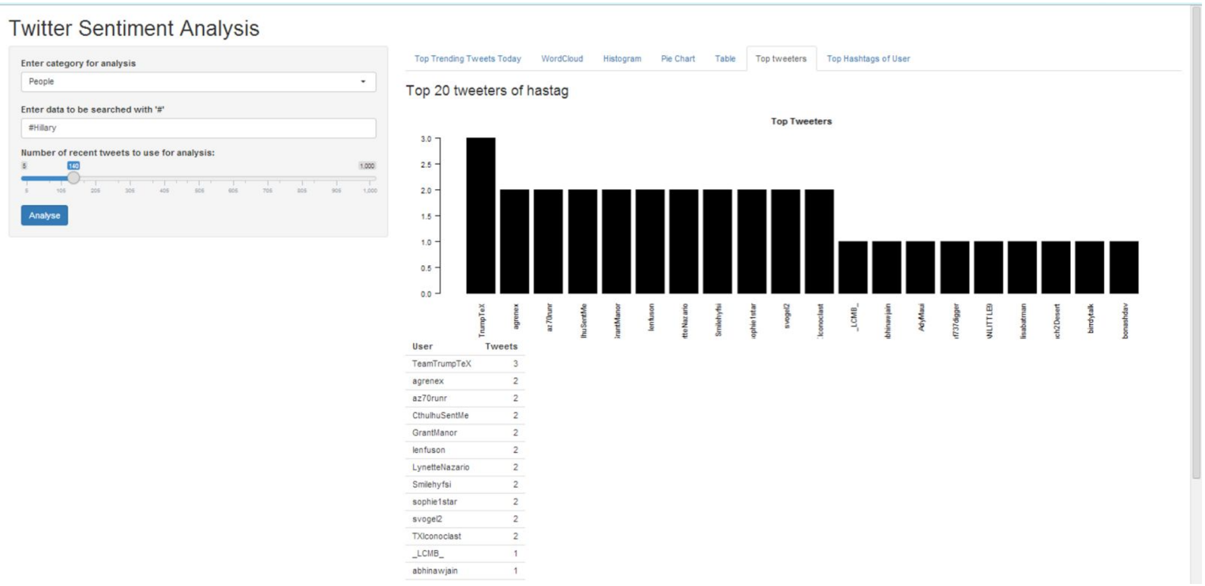
Number of recent tweets to use for analysis: 500

Pie chart

A pie chart is a circular statistical graphic, which is divided into slices to illustrate the sentiment of the hashtag. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.

10. Top Tweeters of Hashtag tab : barplot

Top tweeters of the given hashtag is shown as a bar plot according to the frequency with which they used the hashtag in the last 7 days. Following this is the table of the User Name of tweeter and the frequency for clearer analysis.



Twitter Sentiment Analysis

Enter category for analysis: People

Enter data to be searched with '#': #Hillary

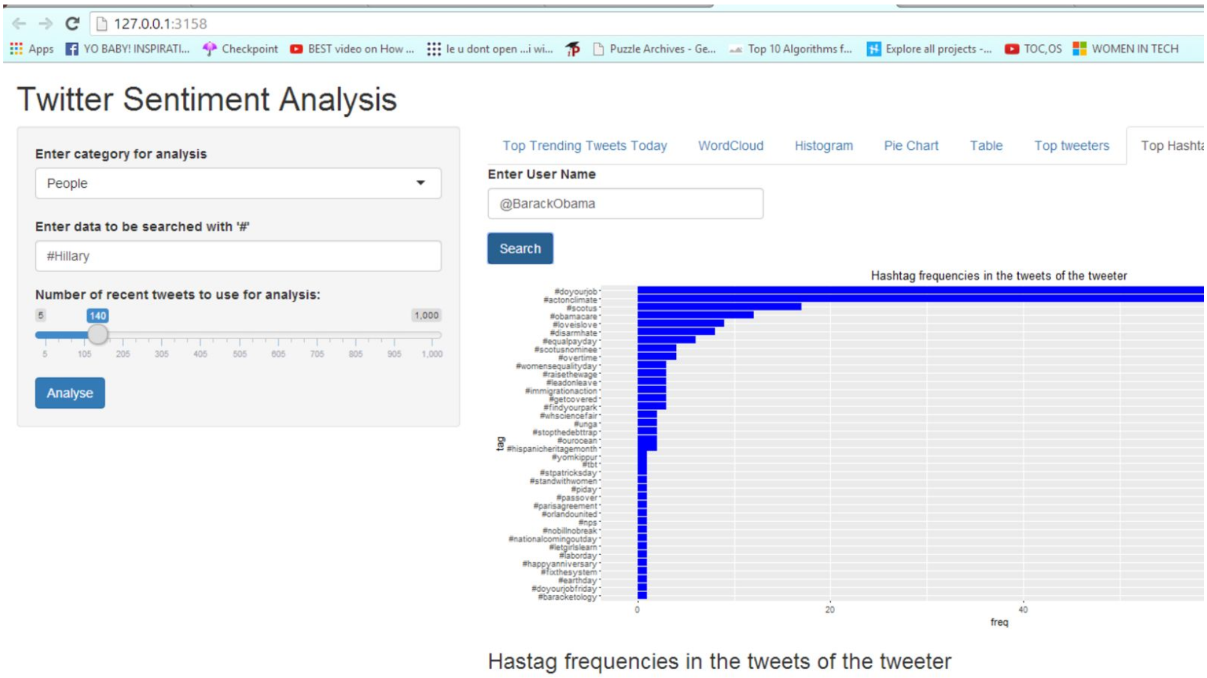
Number of recent tweets to use for analysis: 140

Top 20 tweeters of hastag

User	Tweets
TeamTrumpTeX	3
agrenex	2
az70runr	2
CthulhuSentMe	2
GrantManor	2
lenfuson	2
LynetteNazario	2
Smilehyfsi	2
sophie1star	2
svogel2	2
TXIconoclast	2
LCMB	1
abhinawjain	1

11. Top Hashtags of user : ggplot

The ggplot shows the top hashtags of tweeter along with the frequency of each hashtag. This takes into account the entire user timeline of the tweeter.



Twitter Sentiment Analysis

Enter category for analysis: People

Enter data to be searched with '#': #Hillary

Number of recent tweets to use for analysis: 140

Enter User Name: @BaracKObama

Hastag frequencies in the tweets of the tweeter

CODE

The entire code and the details of each part in a modular version can be find in our Github Repository.There are three sub- repositories: R, Shiny Web Application and Movie Reviews.   The link:

https://github.com/Twitter-Sentiment-Analysis

PACKAGES USED

- **twitteR**: Provides an interface to the Twitter web API
- **stringr**: String operations in R
- **ROAuth**: Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice.
- **RCurl**: Provides functions to allow one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server.
- **ggplot2**: An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources.
- **reshape**: Flexibly restructure and aggregate data using just two functions: melt and cast
- **tm** : A framework for text mining applications within R.
- **RJSONIO**: This is a package that allows conversion to and from data in Javascript object notation (JSON) format. This allows R objects to be inserted into Javascript/ECMAScript/ActionScript code and allows R programmers to read and convert JSON content to R objects
- **wordcloud**: visual representation in the form of wordcloud where size of the word is proportional to the frequency of words used in the tweets
- **gridExtra**: Provides a number of user-level functions to work with "grid" graphics, notably to arrange multiple grid-based plots on a page, and draw tables.
- **plyr**: Tools for Splitting, Applying and Combining Data

LIMITATIONS

1. The Twitter Search API can get tweets upto a maximum of 7 days old.

2. Not effective in detecting sarcasm.

3. Cannot get 100% efficiency in analysing sentiment of tweets.

4. Can only retrieve a maximum of 1000 tweets per query without

authenticating via OAuth before receiving a 403 error or timeout.

5. Giving a hash tag under the wrong category will still give results: No error message

FUTURE WORK

- Detect sarcasm in tweets
- Analyse images for emotions
- Add hindi words to dataset.
- Star rating (Negative and Positive [According to percentage]) (BOX PLOT)
- Find no of mentions of n particular organizations (And analyse sentiment)
- Timeline of 7 days for emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust
- Extract from newspapers(TOI)
- Parallelizing code
- Apply better Machine Learning Algorithms (Like Support Vector Machine)

REFERENCES

- http://www.rdatamining.com/docs/twitter-analysis-with-r
- https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides
- https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
- https://www.quora.com/How-can-I-read-Twitter-data-directly-in-R
- http://www.rdatamining.com/docs/twitter-analysis-with-r
- https://github.com/datumbox/twitter-sentiment-analysis
- https://www.r-bloggers.com/emoticons-decoder-for-social-media-sentiment-analysis-in-r/
- https://www.r-bloggers.com/twitter-sentiment-analysis-with-r/
- https://www.r-bloggers.com/how-to-create-a-twitter-sentiment-analysis-using-r-and-shiny/
- https://www.youtube.com/watch?v=JoArGkOpeU0
- https://eight2late.wordpress.com/2015/11/06/a-gentle-introduction-to-naive-bayes-classification-using-r/
- https://www.youtube.com/watch?v=UwwQrEcXtfc

PROGRESS REPORT

Date	Progress
20.08.16	Basics of R
21.08.16	Created an Application in Twitter and established OAuth to get tweets from Twitter into R
22.08.16	Applied algorithm on real tweets after converting tweets into desired data frame form
23.08.16	Cleaning twitter data, removing emoticons,Improving accuracy function
25.08.16	Configuring Git and Github and uploading files from machine to Github and Github to machine
	Create three data frames:Total, Positive, Negative
26.08.16	Merge the above data frames into one Histograms of Score
	Histograms and Pie charts of Score, Positive, Negative
	Percentage calculation for Positive, Negative
27.08.16	Create pie chart with percentages
	Create pie chart for different degrees of positive, negative and neutral
30.08.16	Research
31.08.16	Removing stop words
	Creating word cloud
01.09.16	Using pre-existing data set to compare sentiment and find accuracy of algorithm using Naive Bayes Classifier
	Implementing algorithm to find sentiment on movie reviews
	Finding top 10 hashtags and their frequency for a tweeter
02.09.16	Finding top ten tweeters (barplot)
	Timeline of particular hashtag (ggplot)
09.09.16	Understanding SVM
	Implementing Bayes(83.2% accuracy achieved!) on movie reviews
11.09.16	Bayes vs Lexical Algorithm Analysis for movie reviews
01.10.16	Basics of Shiny app
02.10.16	Code for Shiny widgets(textInput, radioButtons, sliderInput, selectInput,plot,tabs)
03.10.16	Front End designing completed
04.10.16	Top trending tweets according to location
18.10.16	Debugging, fixing front end errors(name conflict), reactive functions
19.10.16	Debugging, rendering table from backend to front end, authentication error
20.10.16	Integration of front end and back end;
	Extraction of tweets, clean tweets, lexical analysis to calculate score and percentage of each sentiment (Table tab)
21.10.16	Tried to render wordcloud on Shiny (bugs in rendering)
22.10.16	Word cloud bug fixed: context transformation
	Integration of front end and back end for wordcloud, histogram, pie chart and top trending tweets

23.10.16	Made barplot and table for top 20 tweeters for a given hashtag and integrated into front end
24.10.16	Fixing bugs for Reactive environment
25.10.16	Timeline of particular hashtag tab(bug fixing), finishing touches
26.10.16	Top hashtags of given user and integration into front end
27.10.16	Documentation
28.10.16	Documentation