VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM 590014

Data Science with R Project

“Tweet Analysis and Sentiment Mining”

By
Deepthi Bhat (1BM16CS003)
Aditi Awasthi (1BM16CS008)
Medhini Oak (1BM16CS047)

Under the Guidance of
Mr. Vikranth B M
Assistant Professor, Department of CSE
BMS College of Engineering

Data Science with R
Self-study Project carried out at

Department of Computer Science and Engineering
BMS College of Engineering
(Autonomous college under VTU)
P.O. Box No.: 1908, Bull Temple Road, Bangalore-560 019
Aug-Dec 2019

BMS COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Data Science with R titled "**Tweet Analysis and Sentiment Mining**" has been carried out by Deepthi Bhat (1BM16CS003), Aditi Awasthi (1BM16CS008), Medhini Oak (1BM16CS047) during the academic year Aug – Dec 2019.

Signature of the guide
Mr. Vikranth B M
Assistant Professor
Department of Computer Science and Engineering
BMS College of Engineering, Bangalore

BMS COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, Deepthi Bhat (1BM16CS003), Aditi Awasthi (1BM16CS008) and Medhini Oak (1BM16CS047) students of 7th Semester, B.E, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, hereby declare that this Data Science with R project work entitled "**Tweet Analysis and Sentiment Mining**" has been carried out by us under the guidance of Mr. Vikranth B M, Assistant Professor, Department of CSE, BMS College of Engineering, Bangalore during the academic semester Aug-Dec 2019.

We also declare that to the best of our knowledge and belief, the development reported here is not part of any other report by any other students.

Signature

Deepthi Bhat (1BM16CS003)

Aditi Awasthi (1BM16CS008)

Medhini Oak (1BM16CS047)

Abstract

Public and private opinion about a wide variety of subjects is expressed and spread continually via social media. Twitter offers a fast and effective way to analyze users' perspectives. Developing a program for sentiment analysis is an effective computational measure for user perceptions. In this project, we extract the tweets based on their hashtags, analyze the various fields which come along with it and plot various graphs to visually present our findings. Furthermore, we also use Bing and NRC datasets to categorize the tweets into positive or negative (Bing) and in the categories of anger, anticipation, disgust, fear, joy, sadness, surprise or trust (NRC).

Introduction

Sentiment Analysis is the process of computationally determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. With the recent advances in deep learning, the ability of algorithms to analyze text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research.

With more than 321 million active users sending a daily average of 500 million Tweets, Twitter has become one of the top social media platforms for news, information, and interaction with brands and influential figures around the world. Twitter allows businesses to reach a broad audience and connect with customers without intermediaries. Monitoring Twitter allows companies to understand their audience, keep on top of what's being said about their brand and their competitors, and discover new trends in the industry. However, when it comes to analyzing Twitter data, quantitative metrics like the number of mentions or retweets are not enough to get a full picture of a situation. What really counts is being able to grasp the nuance of those mentions. Therefore, Twitter is an ideal platform to perform Sentiment analysis. Some advantages of using Twitter are:

- **Scalability**: The task of extracting a large number of tweets can be automated and cost-effective results can be obtained in a very short time.

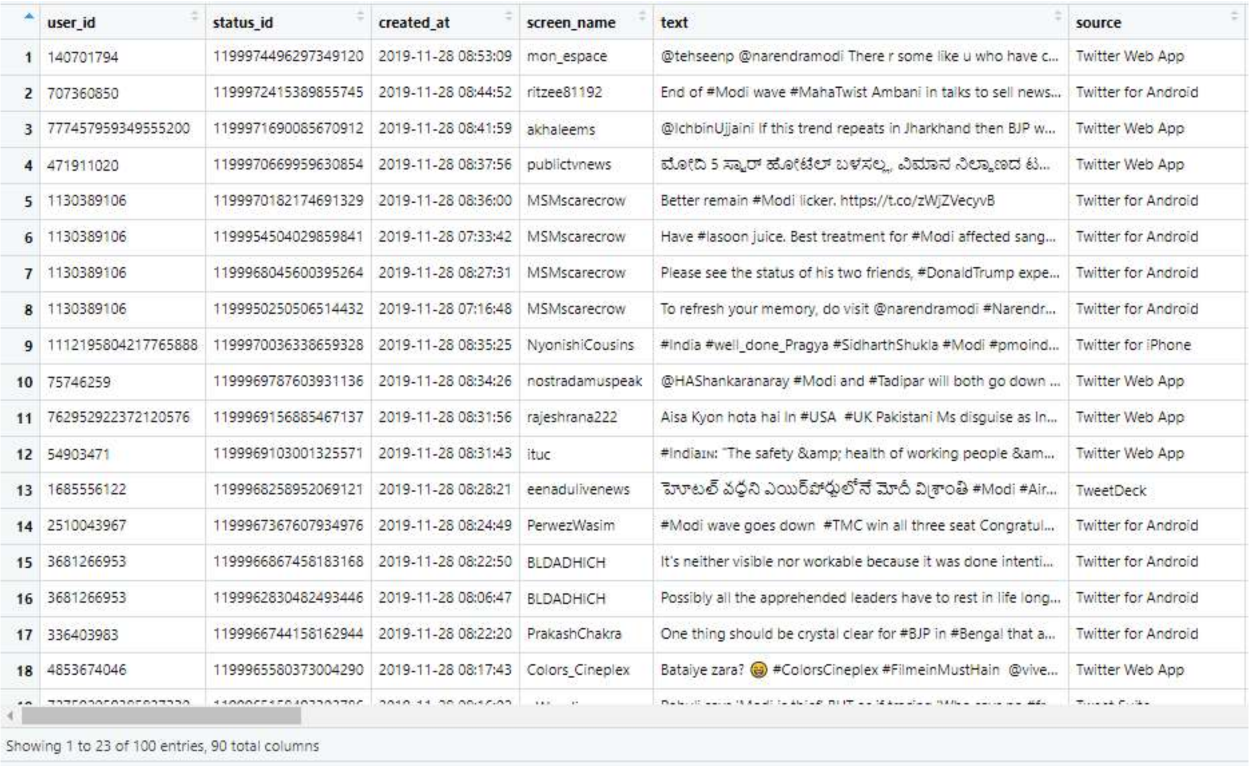
- **Real-Time Analysis:** It is critical to notice sudden shifts in customer moods and detect if critics and complaints are increasing and take action before the problem escalates.

- **Consistent Criteria:** Analyzing sentiment in text is a subjective task. When done manually, the results will probably be biased. Using predefined sentiment datasets help set the parameters to analyze all the data in a uniform fashion and obtain more consistent and accurate results.

Dataset Description

The dataset comprises of tweet information of random users from twitter. To extract this information, Twitter provides its own API. It is necessary to create a developer account and register the application with Twitter. By registering the application, the consumer and access keys are obtained which authenticate the application and enable the extraction of tweets.

The records are extracted as a data frame which comprises of 90 fields like text, location, hashtags and so on.



	user_id	status_id	created_at	screen_name	text	source
1	140701794	1199974496297349120	2019-11-28 08:53:09	mon_espace	@tehseenp @narendramodi There r some like u who have c...	Twitter Web App
2	707360850	1199972415389855745	2019-11-28 08:44:52	ritzee81192	End of #Modi wave #MahaTwist Ambani in talks to sell news...	Twitter for Android
3	777457959349555200	1199971690085670912	2019-11-28 08:41:59	akhaleems	@IchbinUjjaini If this trend repeats in Jharkhand then BJP w...	Twitter Web App
4	471911020	1199970669959630854	2019-11-28 08:37:56	publictvnews	ಮೋದಿ 5 ಸ್ಟಾರ್ ಹೋಟೆಲ್ ಬಳಸಲ್ಲ, ವಿಮಾನ ನಿಲ್ದಾಣದ ಟಿ...	Twitter Web App
5	1130389106	1199970182174691329	2019-11-28 08:36:00	MSMscarecrow	Better remain #Modi licker. https://t.co/zWjZVecyvB	Twitter for Android
6	1130389106	1199954504029859841	2019-11-28 07:33:42	MSMscarecrow	Have #lasoon juice. Best treatment for #Modi affected sang...	Twitter for Android
7	1130389106	1199968045600395264	2019-11-28 08:27:31	MSMscarecrow	Please see the status of his two friends, #DonaldTrump expe...	Twitter for Android
8	1130389106	1199950250506514432	2019-11-28 07:16:48	MSMscarecrow	To refresh your memory, do visit @narendramodi #Narendr...	Twitter for Android
9	1112195804217765888	1199970036338659328	2019-11-28 08:35:25	NyonishiCousins	#India #well_done_Pragya #SidharthShukla #Modi #pmoind...	Twitter for iPhone
10	75746259	1199969787603931136	2019-11-28 08:34:26	nostradamuspeak	@HAShankaranaray #Modi and #Tadipar will both go down ...	Twitter Web App
11	762952922372120576	1199969156885467137	2019-11-28 08:31:56	rajeshrana222	Aisa Kyon hota hai In #USA #UK Pakistani Ms disguise as In...	Twitter Web App
12	54903471	1199969103001325571	2019-11-28 08:31:43	ituc	#Indiaɪɴ: "The safety &amp; health of working people &am...	Twitter Web App
13	1685556122	1199968258952069121	2019-11-28 08:28:21	eenadulivenews	హోటల్ వద్దని ఎయిర్‌పోర్టులోనే మోదీ విశ్రాంతి #Modi #Air...	TweetDeck
14	2510043967	1199967367607934976	2019-11-28 08:24:49	PerwezWasim	#Modi wave goes down #TMC win all three seat Congratul...	Twitter for Android
15	3681266953	1199966867458183168	2019-11-28 08:22:50	BLDADHICH	It's neither visible nor workable because it was done intenti...	Twitter for Android
16	3681266953	1199962830482493446	2019-11-28 08:06:47	BLDADHICH	Possibly all the apprehended leaders have to rest in life long...	Twitter for Android
17	336403983	1199966744158162944	2019-11-28 08:22:20	PrakashChakra	One thing should be crystal clear for #BJP in #Bengal that a...	Twitter for Android
18	4853674046	1199965580373004290	2019-11-28 08:17:43	Colors_Cineplex	Bataiye zara? 😆 #ColorsCineplex #FilmeinMustHain @vive...	Twitter Web App

Showing 1 to 23 of 100 entries, 90 total columns

The fields of the dataset include:

```
 "user_id"                  "status_id"                "created_at"               "screen_name"
 "text"                     "source"                   "display_text_width"       "reply_to_status_id"
 "reply_to_user_id"         "reply_to_screen_name"     "is_quote"                 "is_retweet"
 "favorite_count"           "retweet_count"            "quote_count"              "reply_count"
 "hashtags"                 "symbols"                  "urls_url"                 "urls_t.co"
 "urls_expanded_url"        "media_url"                "media_t.co"               "media_expanded_url"
 "media_type"               "ext_media_url"            "ext_media_t.co"           "ext_media_expanded_url"
 "ext_media_type"           "mentions_user_id"         "mentions_screen_name"     "lang"
 "quoted_status_id"         "quoted_text"              "quoted_created_at"        "quoted_source"
 "quoted_favorite_count"    "quoted_retweet_count"     "quoted_user_id"           "quoted_screen_name"
 "quoted_name"              "quoted_followers_count"   "quoted_friends_count"     "quoted_statuses_count"
 "quoted_location"          "quoted_description"       "quoted_verified"          "retweet_status_id"
 "retweet_text"             "retweet_created_at"       "retweet_source"           "retweet_favorite_count"
 "retweet_retweet_count"    "retweet_user_id"          "retweet_screen_name"      "retweet_name"
 "retweet_followers_count"  "retweet_friends_count"    "retweet_statuses_count"   "retweet_location"
 "retweet_description"      "retweet_verified"         "place_url"                "place_name"
 "place_full_name"          "place_type"               "country"                  "country_code"
 "geo_coords"               "coords_coords"            "bbox_coords"              "status_url"
 "name"                     "location"                 "description"              "url"
 "protected"                "followers_count"          "friends_count"            "listed_count"
 "statuses_count"           "favourites_count"         "account_created_at"       "verified"
 "profile_url"              "profile_expanded_url"     "account_lang"             "profile_banner_url"
 "profile_background_url"   "profile_image_url"
```

Libraries used

The following are the libraries and other requirements needed to run the project.

- **rtweet**: An implementation of calls designed to extract and organize Twitter data via Twitter's REST and stream APIs

- **ggplot2**: A system for 'declaratively' creating graphics and elegant data visualizations

- **dplyr**: A fast, consistent tool for working with data frame like objects, both in memory and out of memory

- **tidytext**: Text mining for word processing and sentiment analysis using 'dplyr', 'ggplot2', and other tidy tools. It provides functions for Bing sentiment analysis and NRC sentiment analysis

- **devtools**: The aim of devtools is to make package development easier by providing R functions that simplify and expedite common tasks

- **widyr**: Encapsulates the pattern of untidying data into a wide matrix, performing some processing, then turning it back into a tidy form

- **tidyr**: The goal of tidyr is to help you create tidy data. Tidy data describes a standard way of storing data that is used wherever possible throughout the tidyverse

- **igraph**: Routines for simple graphs and network analysis. It can handle large graphs very well and provides functions for generating random and regular graphs, graph visualization, centrality methods and much more

- **ggraph**: ggraph is an extension of the ggplot2 API tailored to graph visualizations and provides the same flexible approach to building up plots layer by layer

- **rjson**: Converts R object into JSON objects and vice-versa

- **httr**: Useful tools for working with HTTP organized by HTTP verbs (GET(), POST(),etc.)

- **leaflet**: Create and customize interactive maps using the 'Leaflet' JavaScript library and the 'htmlwidgets' package

- **lubridate** : Provides functions to work with date-times and time-spans

- **zoo**: A class with methods for totally ordered indexed observations

Functionalities

A. Extracting the tweets based on hashtag

In order to fetch tweets through Twitter API, one needs to register an app through their twitter account. The tweets are extracted using the search_tweets() function in the rtweet library after using an authorized consumer key. Users are extracted using search_user().

B. Plotting frequency of tweets location wise and geocoding

By plotting the tweets based on user location, the interest of users based on geographical location can be evaluated. These locations are marked on the map.

C. Getting the frequency of tweets in the past 9 days

Find the extent to which the topic has been trending over the past 9 days.

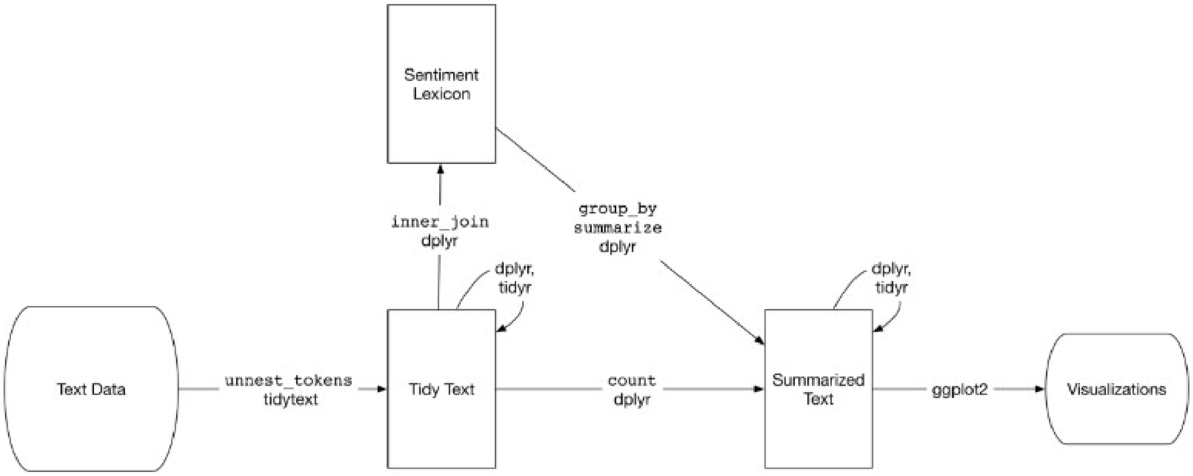
D. Plotting the frequency of tweets by two different news hubs

This shows number of tweets by different media outlets aggregated by hour.

E. Building a Word Network

Before performing sentiment mining, tweets are analyzed to understand the relationships between words. For this, a word network is created which groups similarly occurring words together.

F. Sentiment Analysis



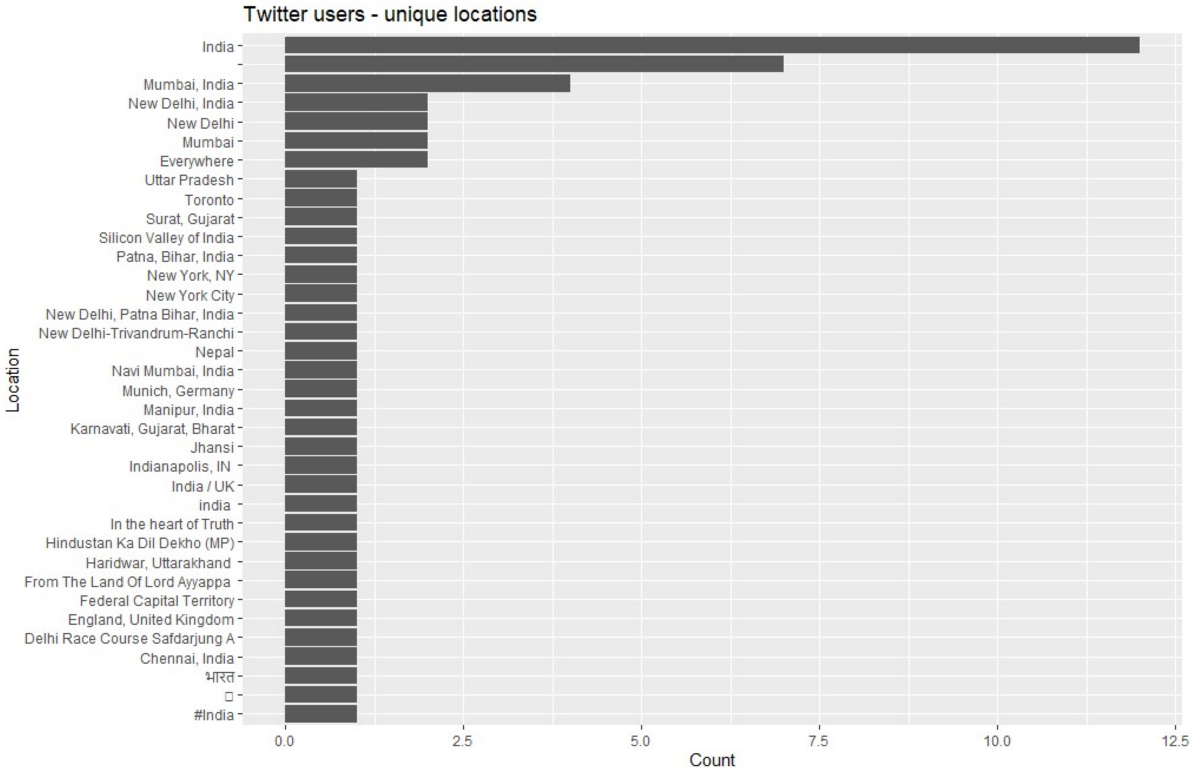
a) **Bing Dataset**

It categorizes words in a binary fashion into positive and negative categories.

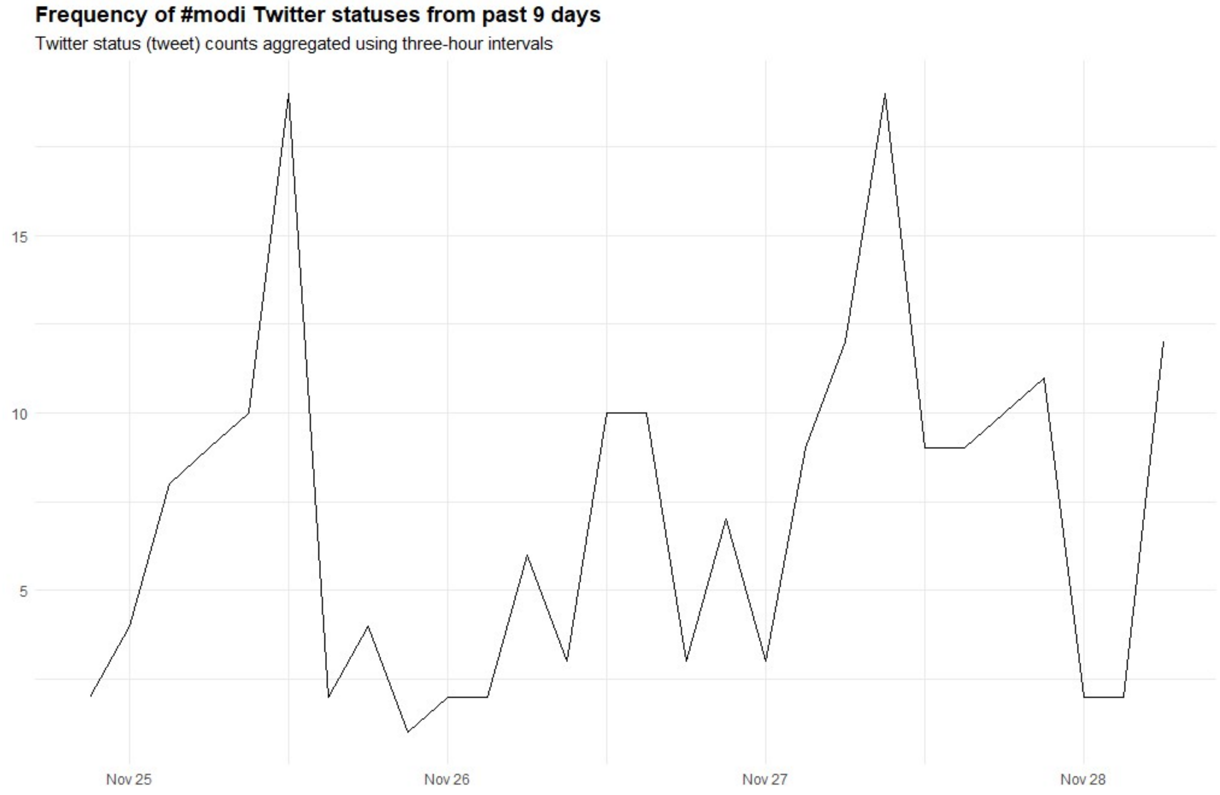
b) **NRC Dataset**

Just classifying the tweets into positive and negative may not give a complete understanding of the sentiment. NRC dataset further helps to categorize the sentiments into anger, anticipation, disgust, fear, joy, sadness, surprise or trust.

Visualization

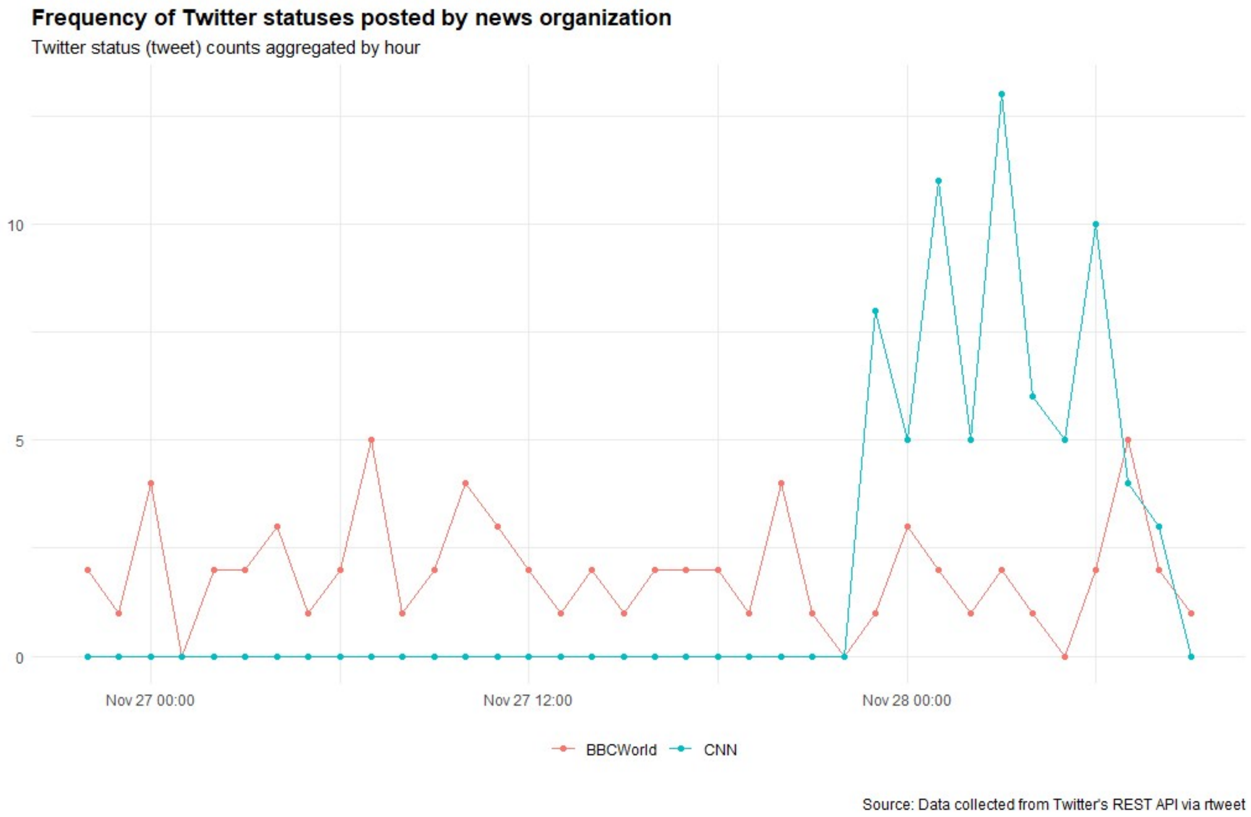


Twitter users - unique locations



Frequency of #modi Twitter statuses from past 9 days

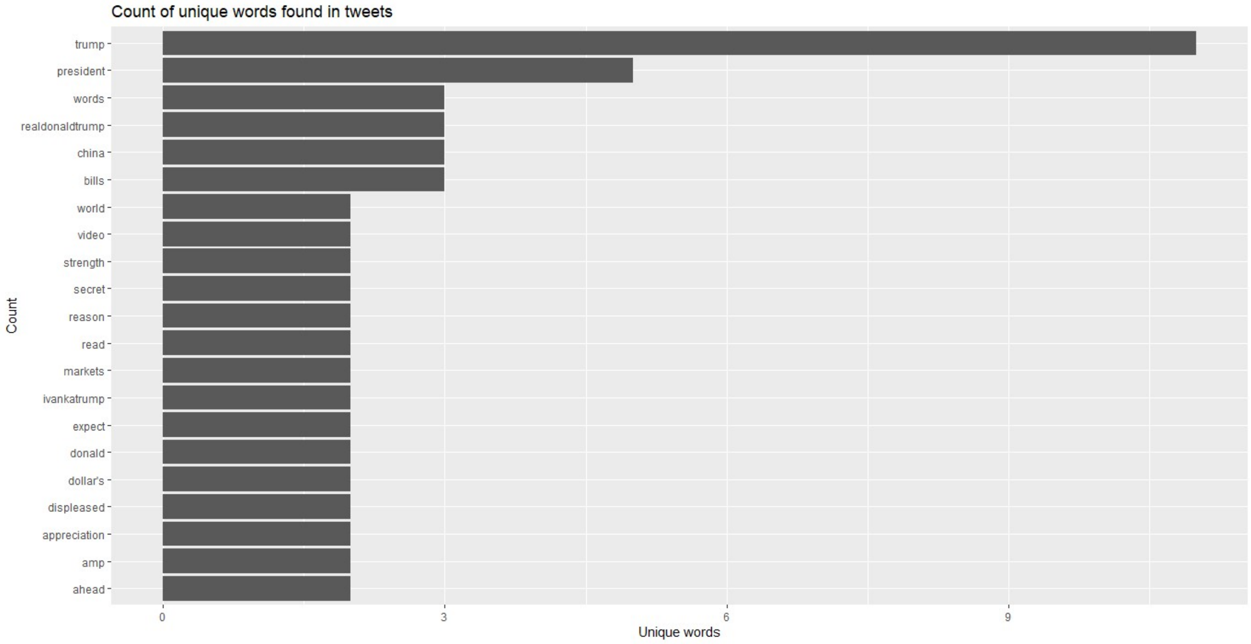
Twitter status (tweet) counts aggregated using three-hour intervals



Frequency of Twitter statuses posted by news organization

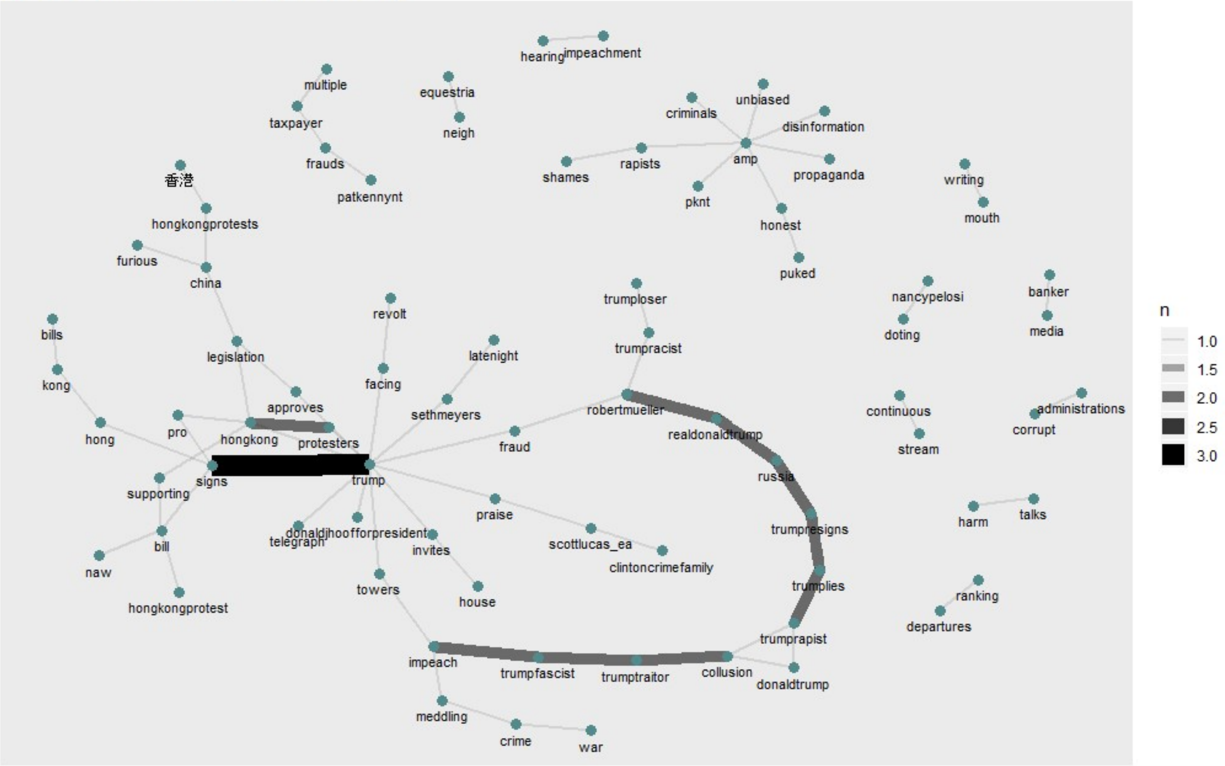
Twitter status (tweet) counts aggregated by hour

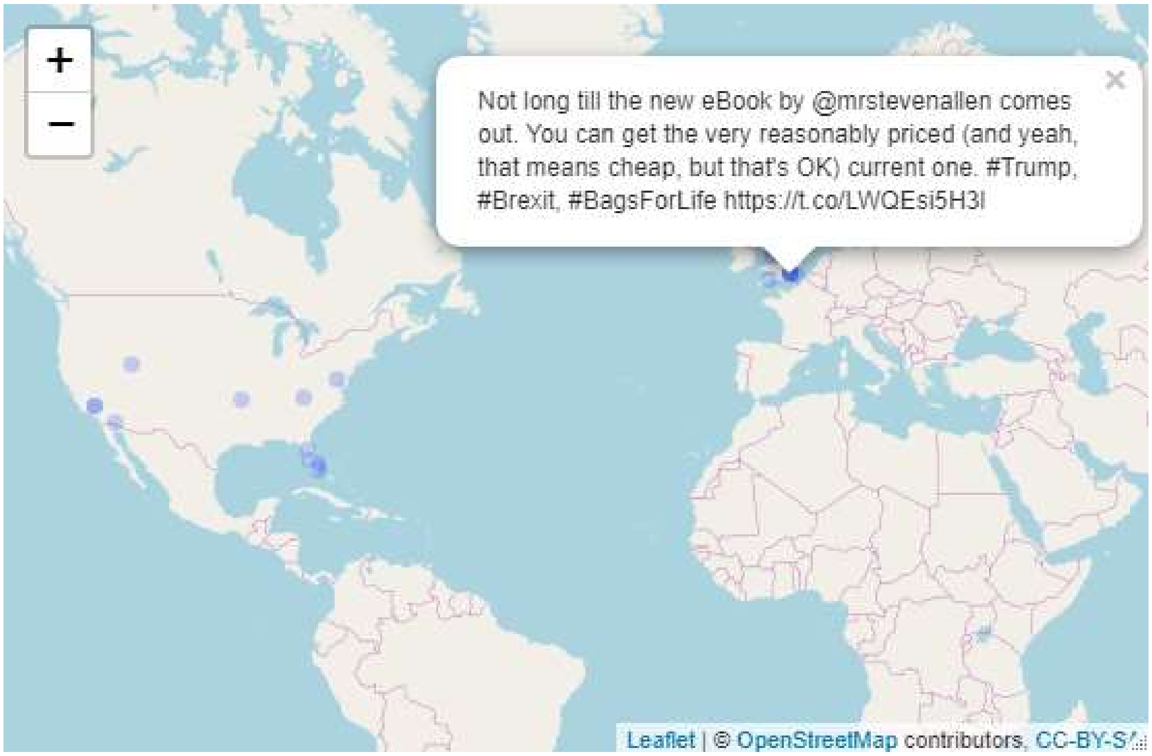
Source: Data collected from Twitter's REST API via rtweet



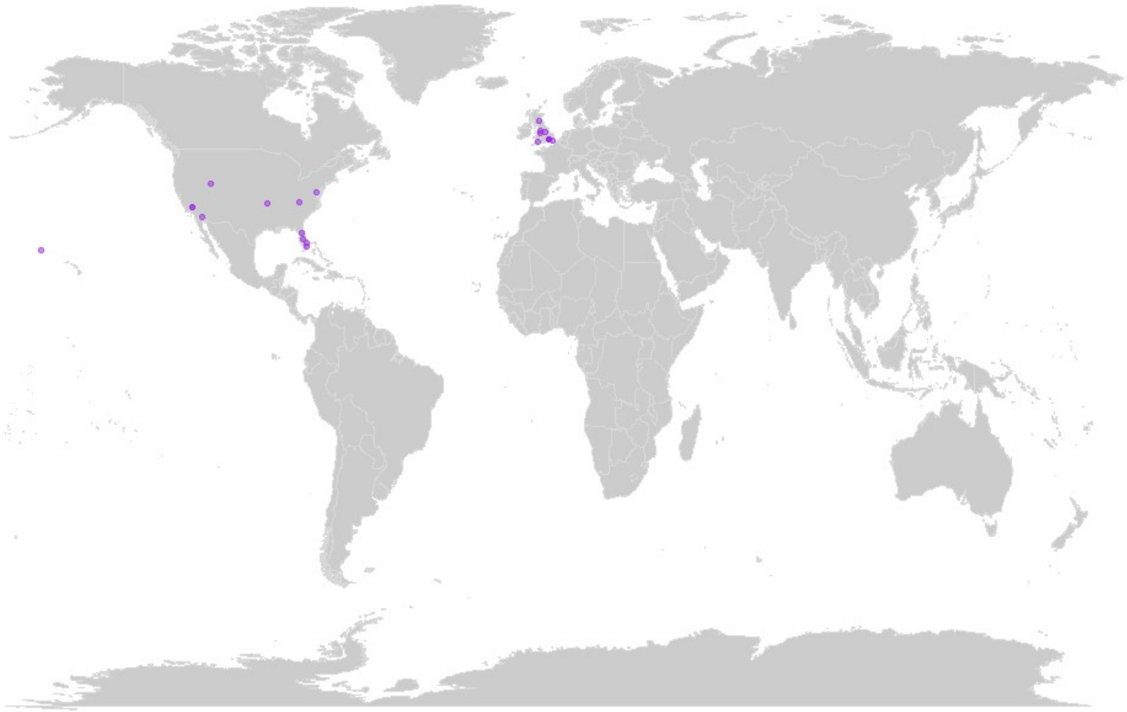
Count of unique words found in tweets

Word Network: Tweets using the hashtag

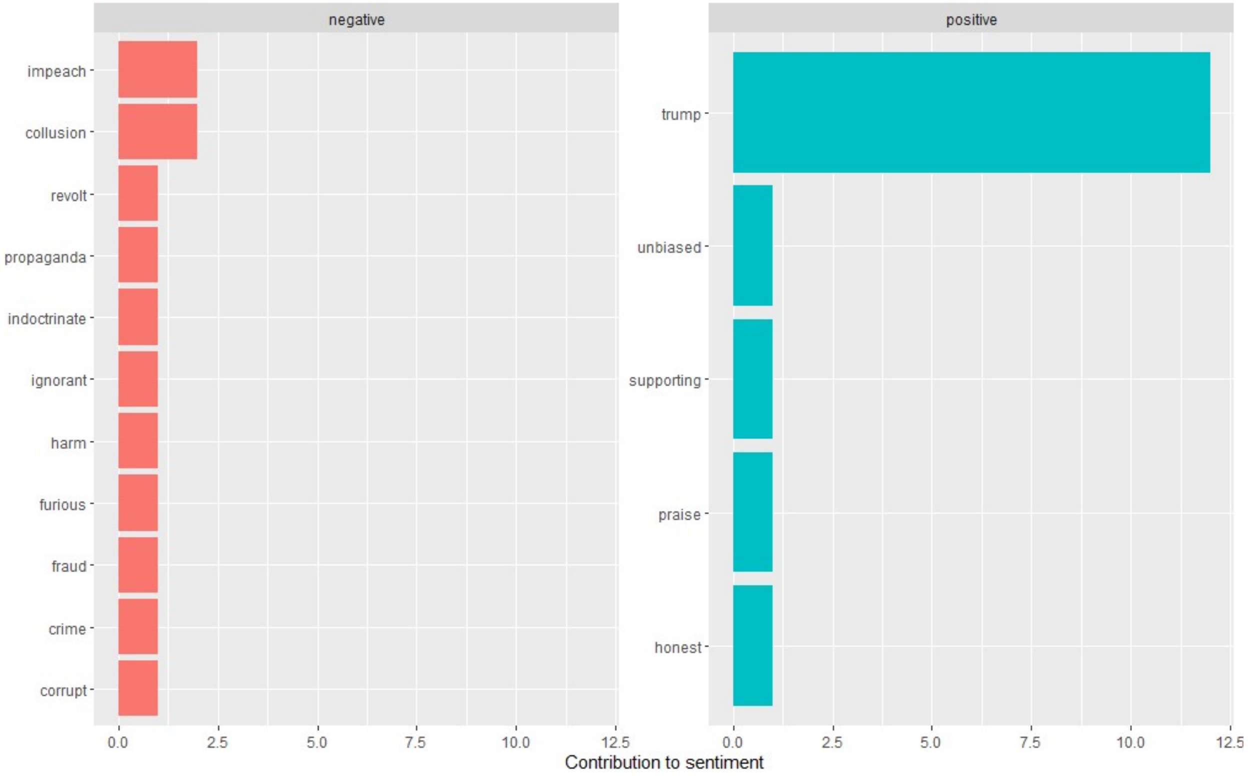




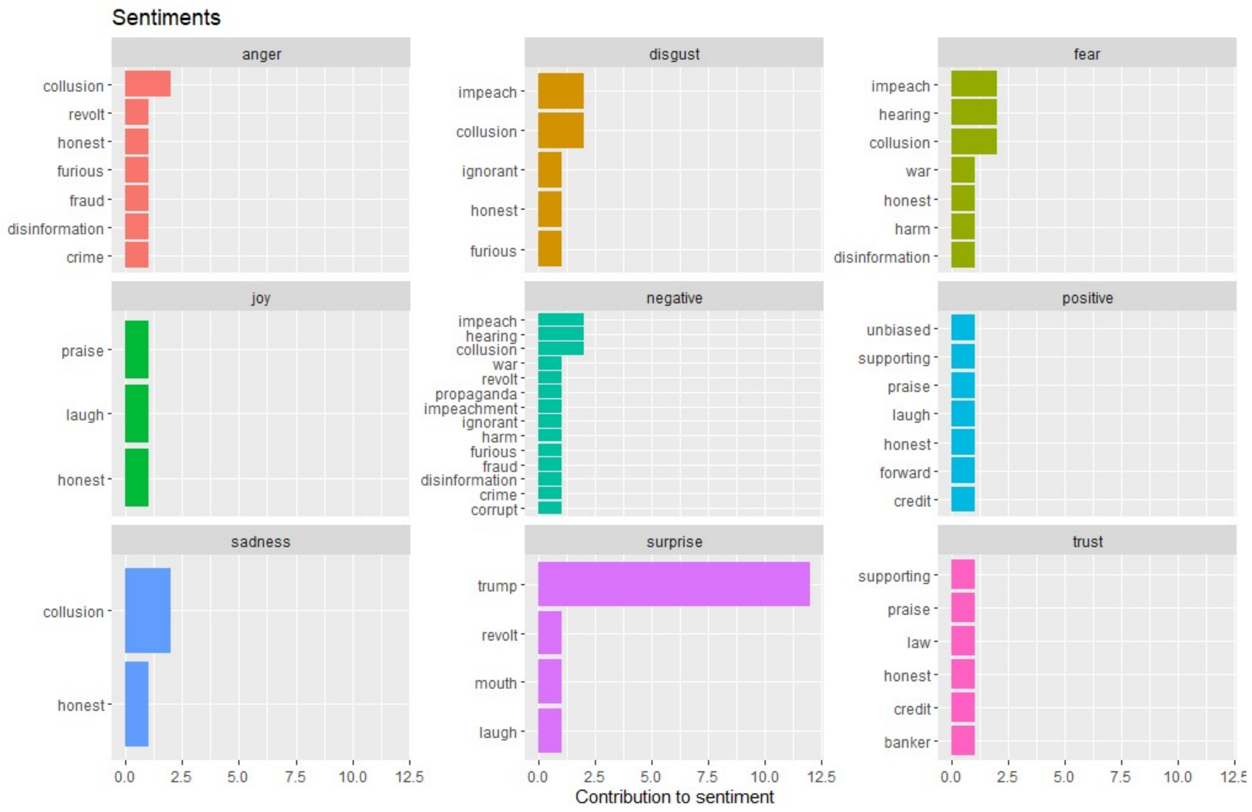
Not long till the new eBook by @mrstevenallen comes out. You can get the very reasonably priced (and yeah, that means cheap, but that's OK) current one. #Trump, #Brexit, #BagsForLife https://t.co/LWQEsi5H3l

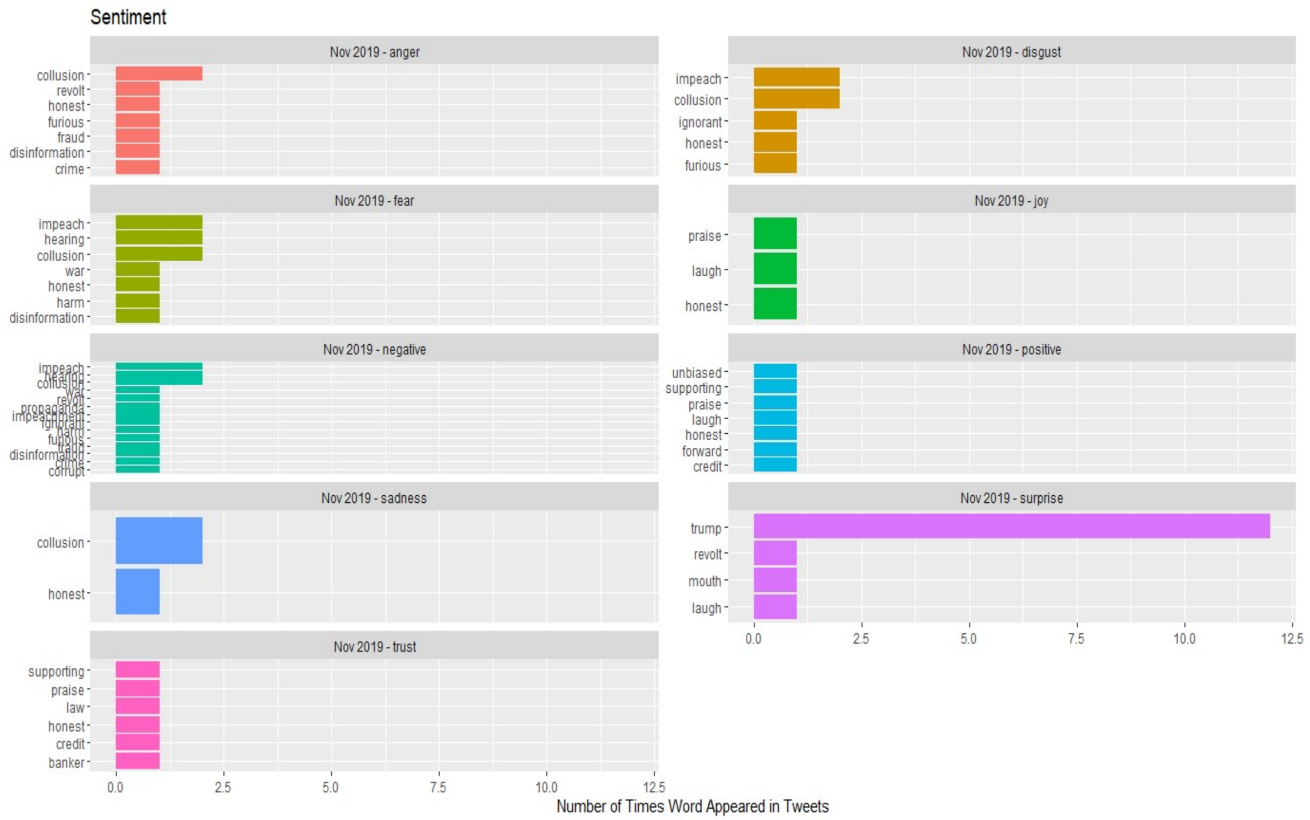


Tweet Locations



Sentiments


Sentiments
anger
collusion
revolt
honest
furious
fraud
disinformation
crime
disgust
impeach
collusion
ignorant
honest
furious
fear
impeach
hearing
collusion
war
honest
harm
disinformation
joy
praise
laugh
honest
negative
impeach
hearing
collusion
war
revolt
propaganda
impeachment
ignorant
harm
furious
fraud
disinformation
crime
corrupt
positive
unbiased
supporting
praise
laugh
honest
forward
credit
sadness
collusion
honest
surprise
trump
revolt
mouth
laugh
trust
supporting
praise
law
honest
credit
banker
0.0
2.5
5.0
7.5
10.0
12.5
Contribution to sentiment


Sentiment
Nov 2019 - anger
collusion
revolt
honest
furious
fraud
disinformation
crime
Nov 2019 - disgust
impeach
collusion
ignorant
honest
furious
Nov 2019 - fear
impeach
hearing
collusion
war
honest
harm
disinformation
Nov 2019 - joy
praise
laugh
honest
Nov 2019 - negative
impeach
hearing
collusion
war
revolt
propaganda
impeachment
ignorant
harm
furious
fraud
disinformation
crime
corrupt
Nov 2019 - positive
unbiased
supporting
praise
laugh
honest
forward
credit
Nov 2019 - sadness
collusion
honest
Nov 2019 - surprise
trump
revolt
mouth
laugh
Nov 2019 - trust
supporting
praise
law
honest
credit
banker
0.0
2.5
5.0
7.5
10.0
12.5
Number of Times Word Appeared in Tweets

Conclusion

This project can provide accurate public opinion about various socially relevant topics just by analyzing tweets from Twitter. It is also enables the user to visualize the results with the help of graphs, frequency charts and other visual aids. This project can be applied in various fields to draw appropriate conclusions. Some of them are:

- **Business:** In the field of marketing, companies use it to develop their strategies, to understand customers' feelings towards products or brand and how people respond to their campaigns or product launches.

- **Politics:** In the political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well.

- **Public Actions:** Sentiment analysis also is used to monitor and analyze social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.