|  |
| --- |
|  |
| CP G1 Group Assignment |
| Web & Social Media Analytics – Brand perception analysis (Text mining) |



Amit Sharma

Amitesh Bajpai

Sahil Sachdeva

Kushal Maheshwari

Neeraj Singh

PROBLEM STATEMENT

Objective: Brand perception analysis (Social media analytics - Text mining)

* Identify a brand - any global or Indian product, celebrity, company, etc.
* For the identified brand download minimum 1000 twitter messages for the most recent period.
* Perform EDA and Data Cleaning.
* Conduct text mining on the data – Correlation, Frequency, Topic Modelling using Word Association, Sentiment Analysis.

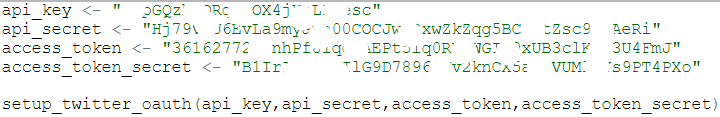
TOPIC SELECTION

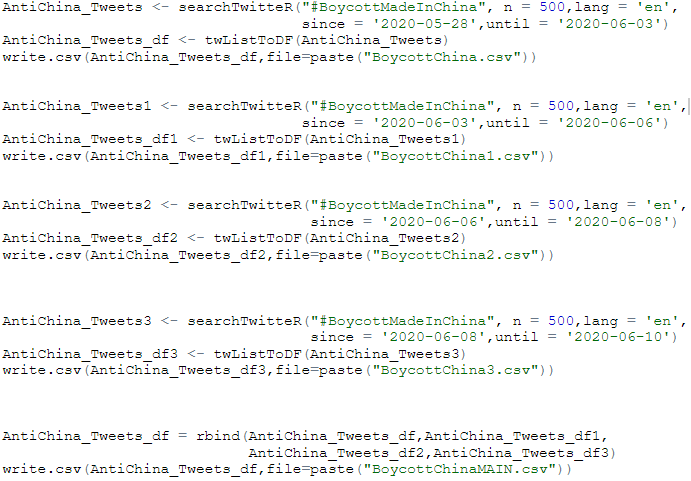
With the Emergence of social media high quality of structured and unstructured information shared through various sources such as the data generated by Twitter or Facebook which depicts user sentiments. The ability to process this information has become important to deep dive into the Brand Perception using Twitter Sentiment Analysis. Text analysis is a part of machine learning technique where the high-quality information derived from text to mine the customer perception about a particular brand.

Brand perception is a special result of a consumer’s experiences with a brand. The purpose of our analysis is to find awareness and reaction of Indians related to a trend ***#BoycottMadeInChina*** which has been started by a renowned personality Mr. Sonam Wangchuk via his YouTube video on 28th May. This report tries to do text mining and sentiment analysis of reactions for last 10-12 days (28 May – 10 June). For this we collected data(tweets) from twitter.

SCRAPPING OF TWEETS

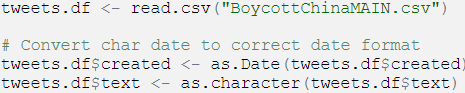
We extracted tweets for 4 different periods related to trend ***#BoycottMadeInChina*** getting around 500 tweets per period and combined them into one dataset. We used twitter developer account for scrapping.





EXPLORATORY DATA ANALYSIS

We read the dataset into R. Before cleaning the data for analysis, let’s first change the format of date and text columns. Text column might contain numbers, hash tags and other regular expressions converting them all to character value making easy for corpus to read.



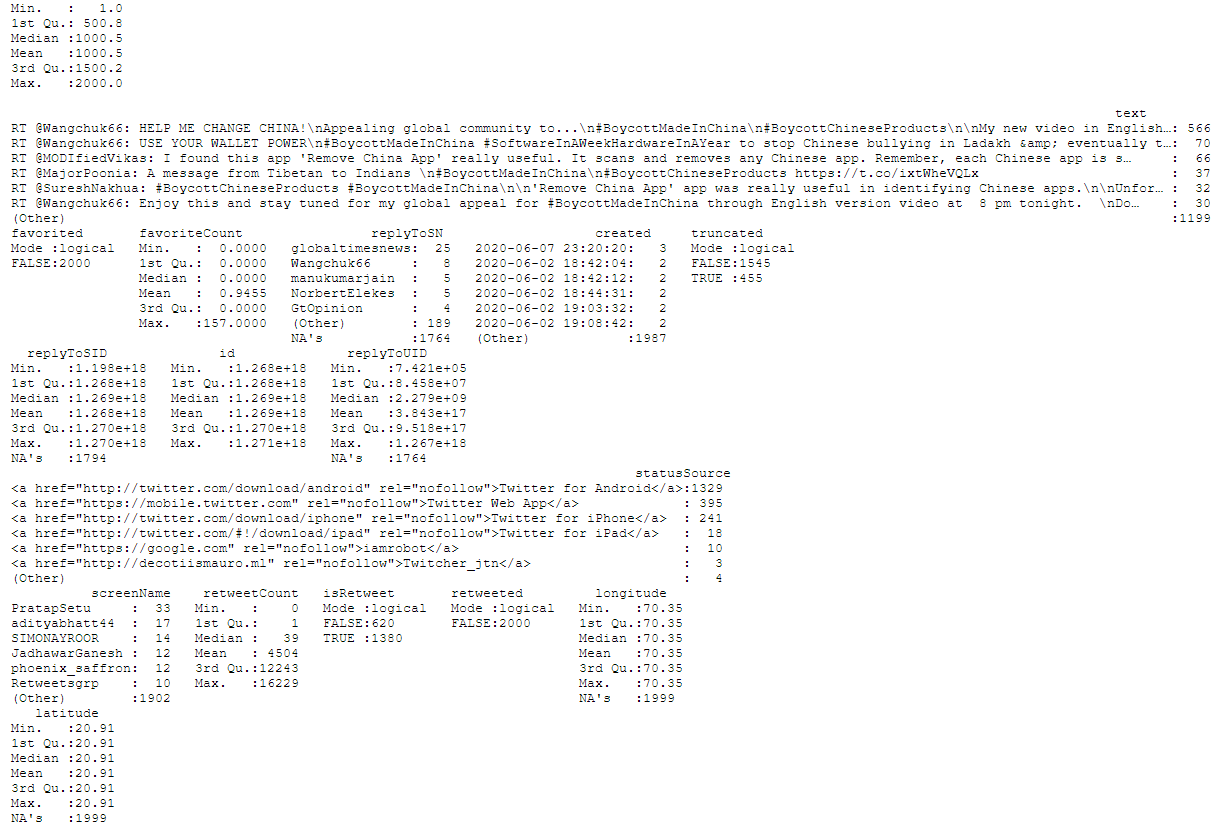
Let’s check the structure of dataset.



Dimensionality of the dataset is given below.



Summarizing the dataset, we get following details.



For the analysis we will be importing following libraries into our R session.



PRE-PROCESSING AND DATA CLEANING

The corpus consists of text document. R programming uses the term to encompass a set of texts considered similar.



The above code reads in the data and returns length of text assigned.

Pre-processing the text can dramatically improve performance of the Bag of Words method. The first step towards doing this is Creating a Corpus, which in simple terms, is nothing but collection of text documents. Once the Corpus created, we are ready for pre-processing.



First, let us remove Punctuation. The basic approach to deal with this is to remove everything that isn’t a standard number or letter. In our case, we will remove all punctuation.

Next, we change the case of the word to lowercase so that same words are not counted as different because of lower or upper case.

Another pre-processing task we must do is to remove meaningless terms to improve our ability to understand sentiments. Transformations in text done via the **tm\_map()** function. Basically, all transformations work on single text documents and **tm\_map()** just applies them to all documents in a corpus.

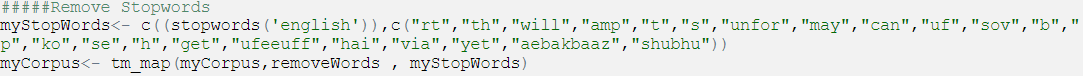
*Convert text to lowercase*

The process of normalization involves transforming text uniformly.



*Remove stop words*

Stop words are just common words which are meaningless. If we look at the result of stop words (“English”) we can see what is getting removed. Let’s create our own stop word removal dictionary to mine text further.



*Remove Numbers*



*Remove Punctuation*



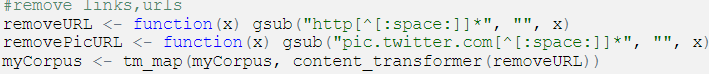
*Strip white spaces*

This will trim white space from the text corpora.



*Remove links/URLs*

This will remove links and URL from the corpus document.



*Remove the @ (usernames)*

Few words in corpus may contain mail id’s or words starting with @.



*Remove anything except the English language (including nos.) and space*



*Remove single letter words*

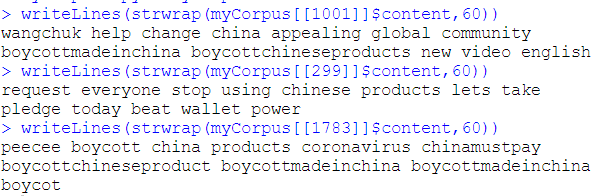


Let’s now look at some tweets in corpus before and after preprocessing task to appreciate the process.

*BEFORE PREPROCESSING*



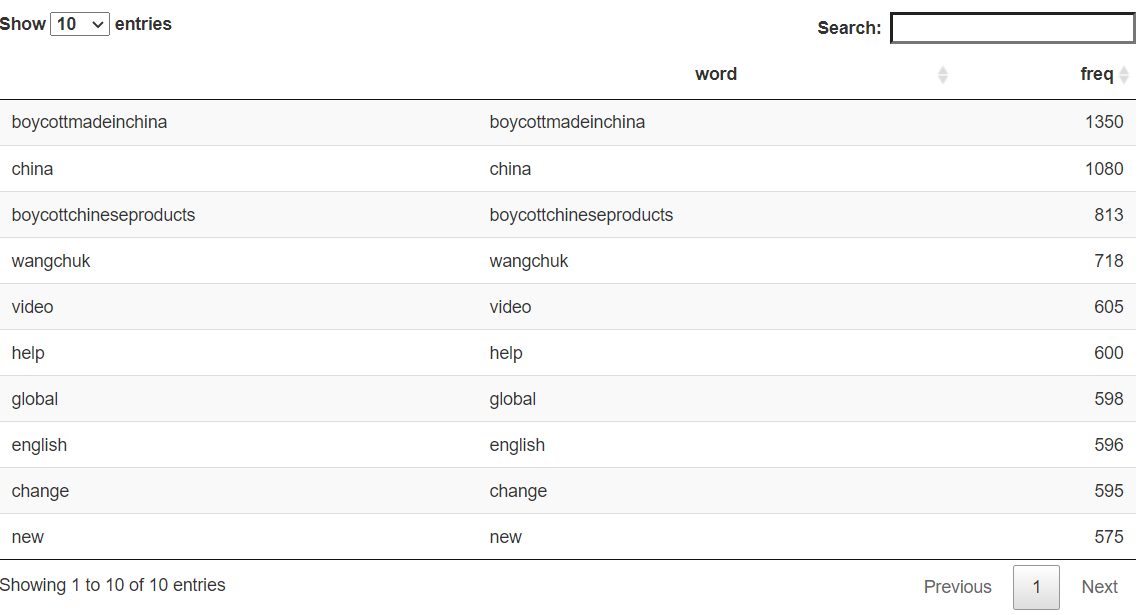
*AFTER PREPROCESSING*

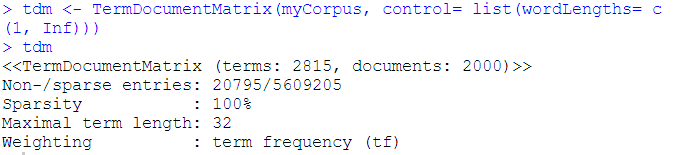


ANALYSING TEXT FREQUENCY

Once we have pre-processed our data, we’re now ready to extract the word frequencies used in our twitter data. The tm package provides a function called Term Document Matrix that generates a matrix where the rows correspond to documents, in our case tweets, and the columns correspond to words in those tweets.

The values in the matrix are the counts of how many times that word appeared in each document. Document matrix is a table containing the frequency of the words. Column names are words and row names are documents.





Words having frequency of at least 25 in the corpus are



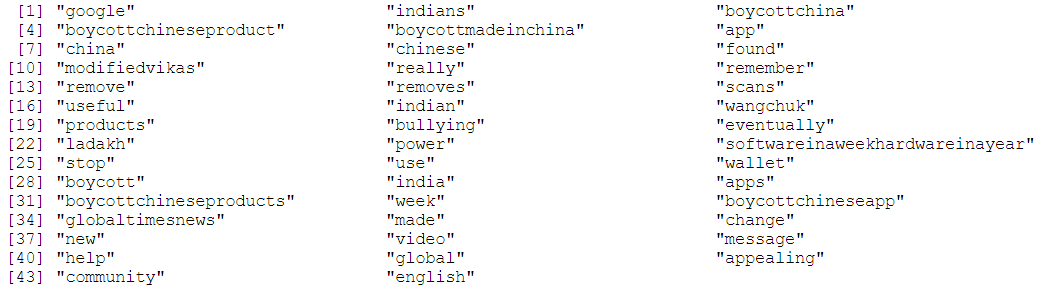
Words having frequency of at least 10 are



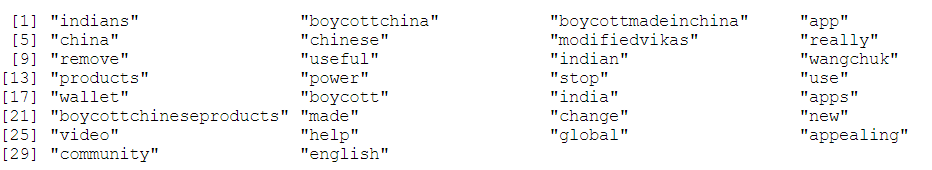




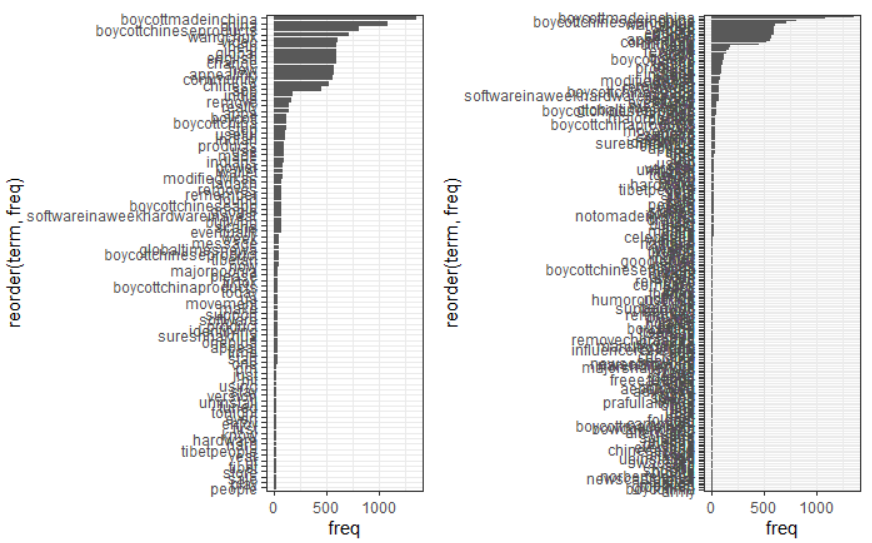
Words having frequency of at least 55 are

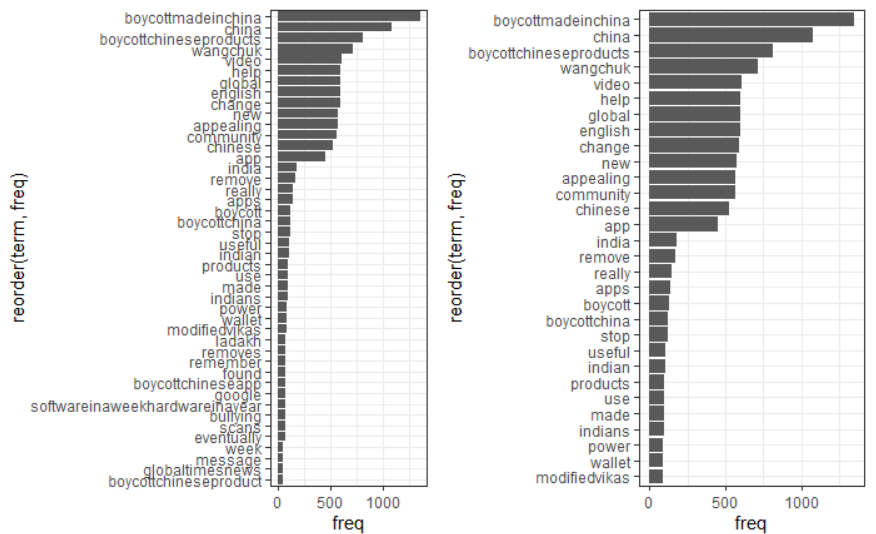


Words having frequency of at least 85 are



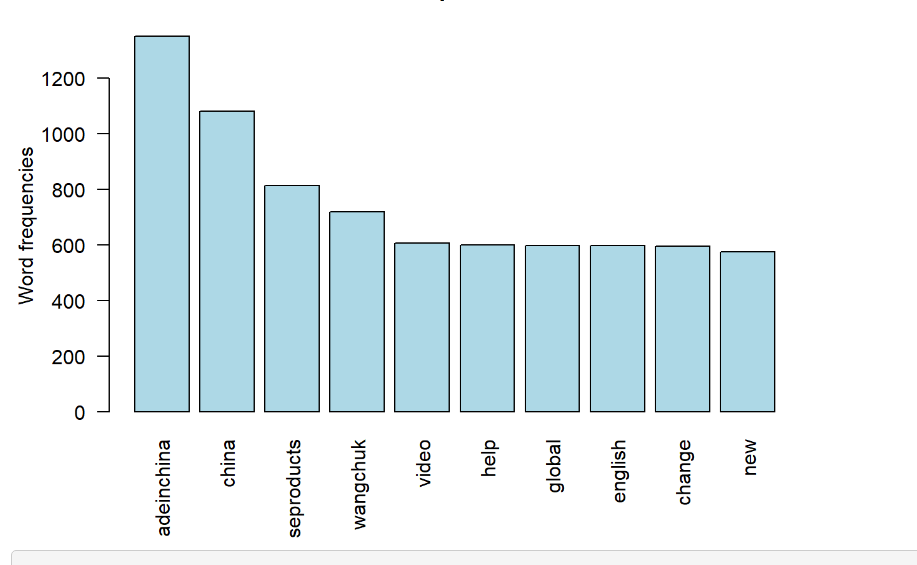
Now let’s plot the graphs of frequent words at 10,25, 55 and 85.





Based on the TermDocumentMatrix() output tried to sort the keywords based on their frequency. The word with high frequency is ***boycottmadeinchina*** as tweeted by the users.

Given below the word frequencies plot.



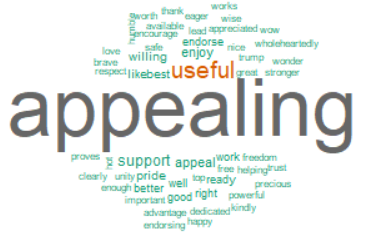
CREATING WORD CLOUD(S)

First of all, we create an overall word cloud followed by a positive only word cloud and negative only word cloud.

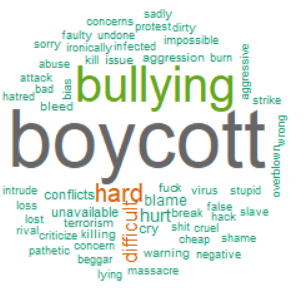


Most frequent words are **boycottmadeinchina, boycottchineseproducts** (both trends), **wangchuk, china, video, change, appealing** etc.

Below is the word cloud of only positive words. Most frequent positive words are **appealing, useful**.

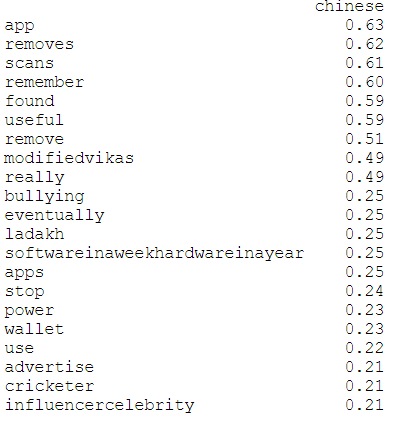


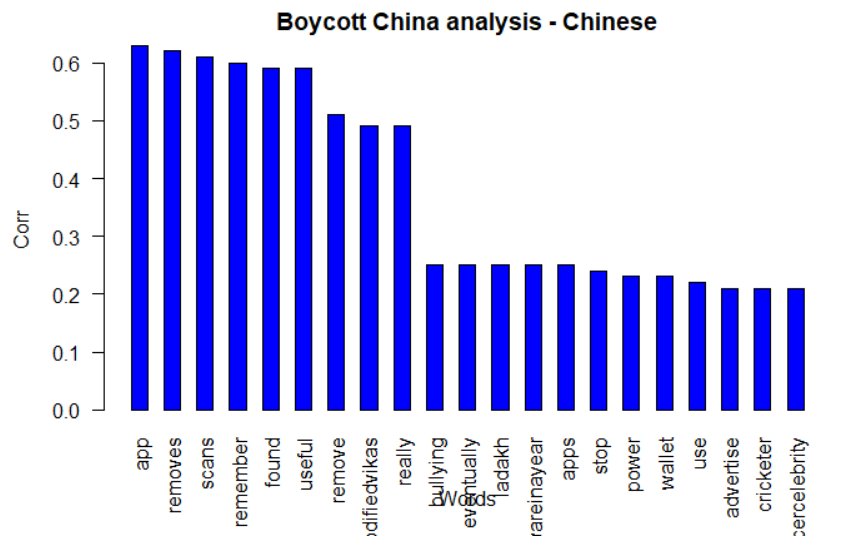
Below is the word cloud of only negative words. Most frequent negative words are **boycott, bullying, hard, difficult**.



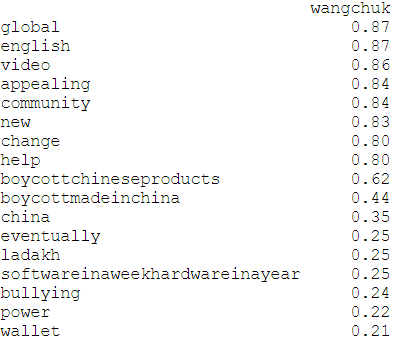
WORD ASSOCIATION AND CORRELATION CHART

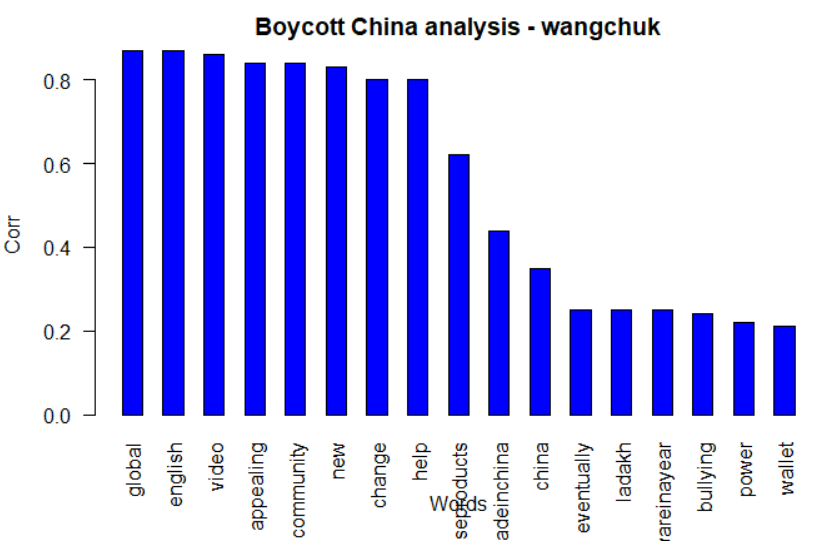
Here we try to find association between the keywords. We have taken words **chinese**, **wangchuk**, **change** and **softwareinaweekhardwareinayear** to find out word association.



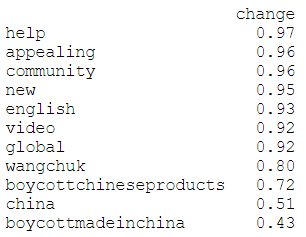


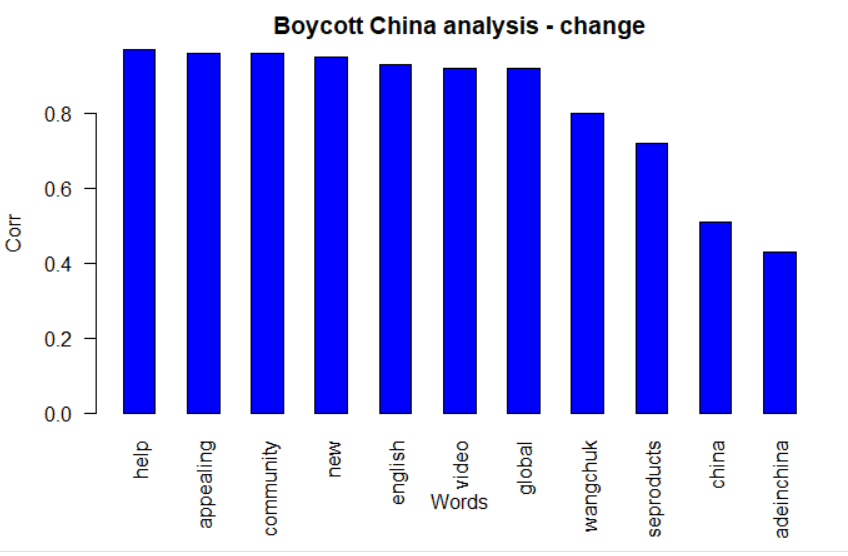
The word **Chinese** is strongly correlated with terms – **app, removes, scans, remember, found, useful** etc.



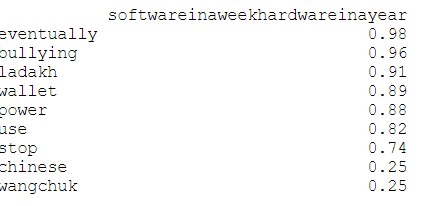


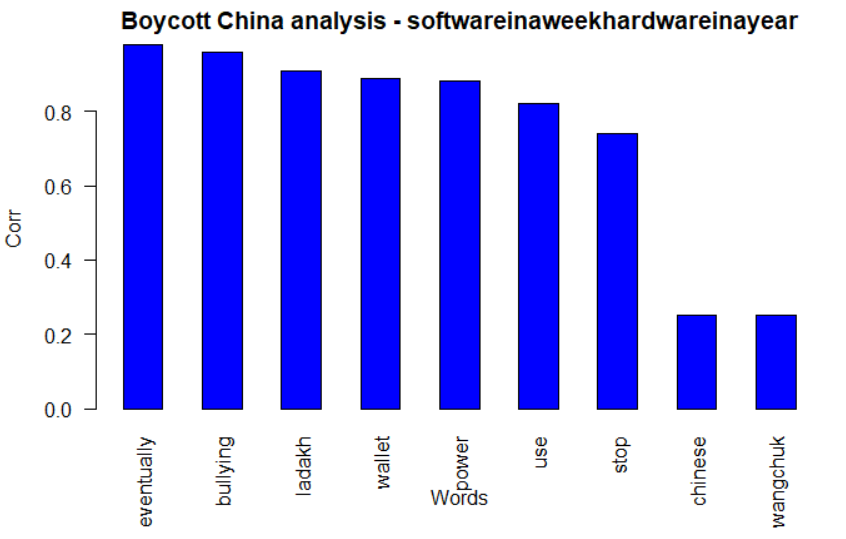
The word **wangchuk** is strongly correlated with terms – **global, English, video, appealing, community, new, change, help** etc.





The word **change** is highly correlated with terms – **help, appealing, community, new, English, video, global** etc.

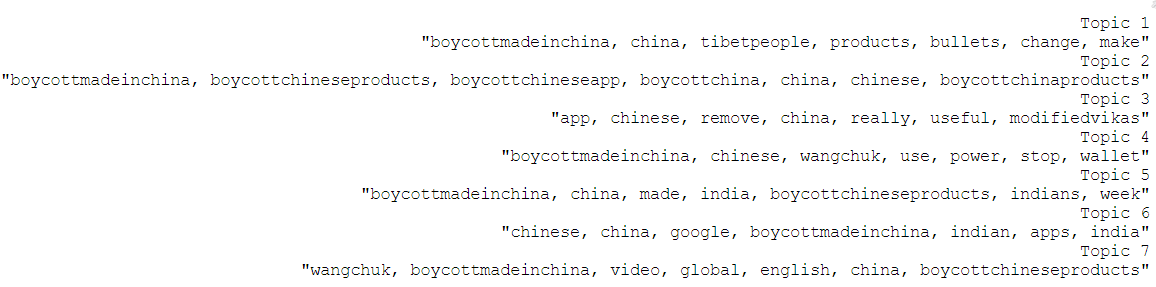




The word **softwareinaweekhardwareinayear** which was a trend started is highly correlated with terms – **eventually, bullying, ladakh, wallet, power** etc.

TOPIC MODELLING USING LDA

We do topic modelling to identify latent/hidden topics using LDA technique. Using this we have found out 7 hidden topics and first 7 terms for these topics.



Topic one says we have to fight china not with bullets but with boycotting their products.

Second topic says about boycotting China and everything Chinese.

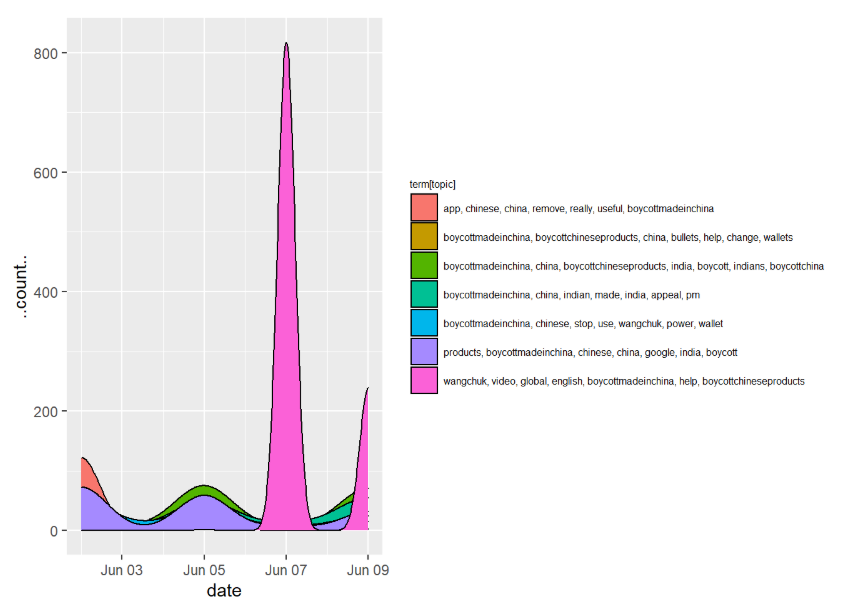
Third topic says we need to remove Chinese apps even though they are useful now.

Fourth topic says we need to use wallet power to stop China.

Fifth topic says we need to use made in India products.

Sixth topic says we ca use Indian alternative apps.

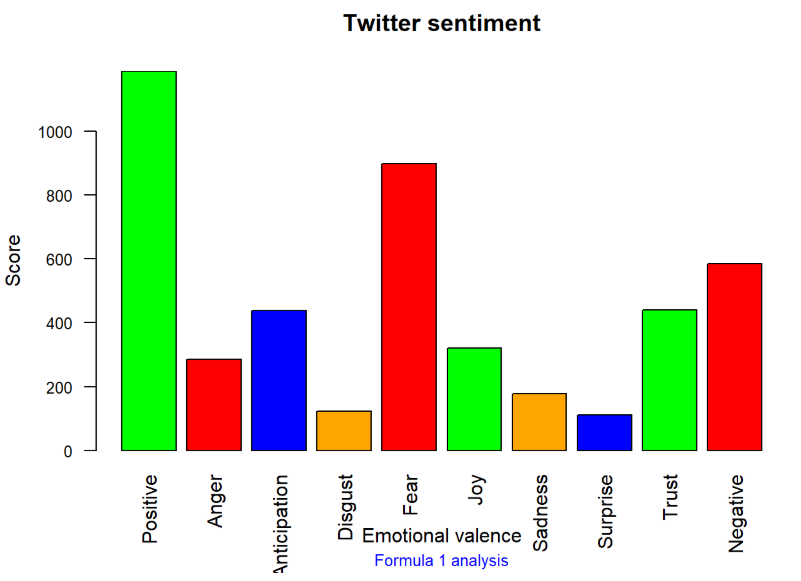
Seventh topic says wangchuk’s appeal of boycotting China to be spread on a global level.



Above plot shows count distribution of topics over 4 days. For example, on June 7 people tweeted mostly about Topic 7.

TWITTER SENTIMENT ANALYSIS

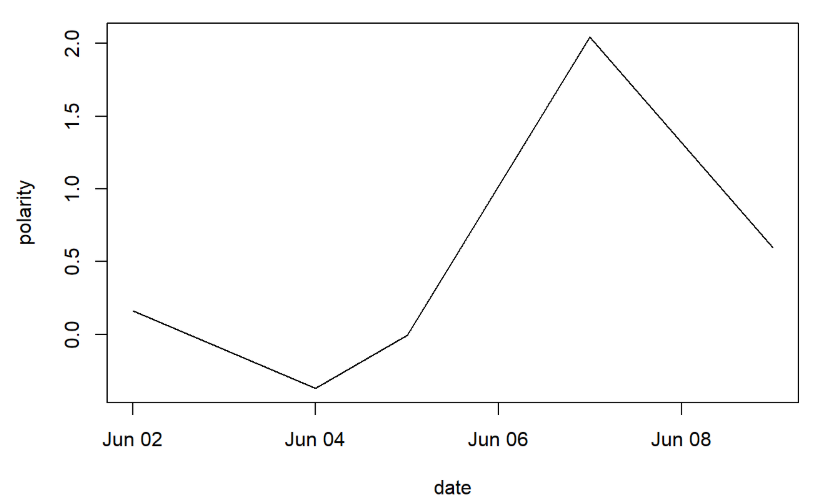
Sentiment analysis and opinion mining is the field of study that analyzes people’s opinions, sentiments, evaluations, attitudes, and emotions from written language. Here, we have sentiments categorized as Positive, Anticipation, Fear, Joy, Surprise and Negative. We understand the emotional valence in tweets.



The score for positive tweets is much higher than any other emotion. This means people in general have reacted positively towards Wangchuk’s appeal of boycotting Chinese goods. Fear also has a good score. This might be because some people think we have over-dependency on Chinese products and boycotting them at once might not be a good idea. Disgust, Sadness and Surprise have lowest scores which is apparent with the context.

POLARITY ANALYSIS

We will try to understand the polarity of tweets on different days and whether it is increasing or decreasing.



The plot show on average, tweets have positive polarity. On June 7 polarity was maximum while on June 2 it was on the negative side.

