**WEB AND SOCIAL MEDIA ANALYTICS ASSIGNMENT**

**BRAND PERCEPTION - CAA**

GROUP 1

## **SOURCE OF DATA**

Data taken for the WMSA Brand Perception Assignment is from **Twitter**.

## **TOPIC SELECTED**

The Subject chosen for this Assignment is **CAA** and the hashtag used for tweet extraction is #CAA (Citizenship Amendment Act).

In our scenario, Excluding Retweets, 1500 English tweets are extracted on the chosen topic. This is done with the help of the “**rtweet**” package which contains the ***search\_tweets()*** function.

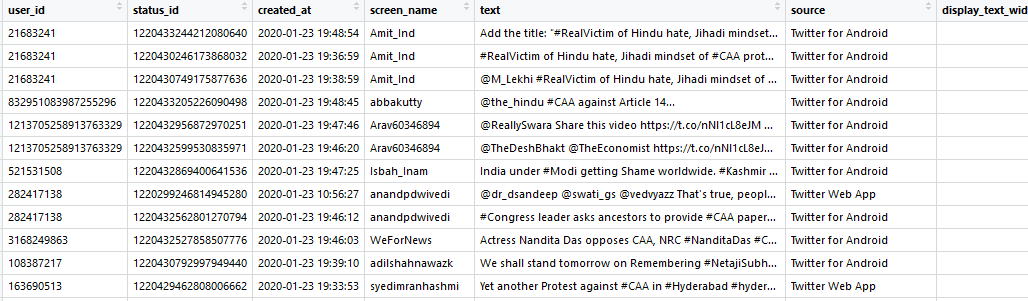


# **EXPLORATORY DATA ANALYSIS**

The data frame structure obtained is viewed. 1499 rows and 90 columns are extracted, wherein column “text” contains the tweets on which further text analytics is performed.



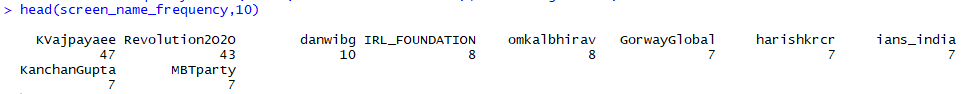
The first few rows and columns extracted are as shown below.



Now the “Screen Names” – display name/author of the tweet and its frequency is taken to understand as to how many users have tweeted how many times.



On an average its seen, most of the users are 1-time tweeters on this topic. Also, the top few screen names who actively posted their views on CAA are found as below.

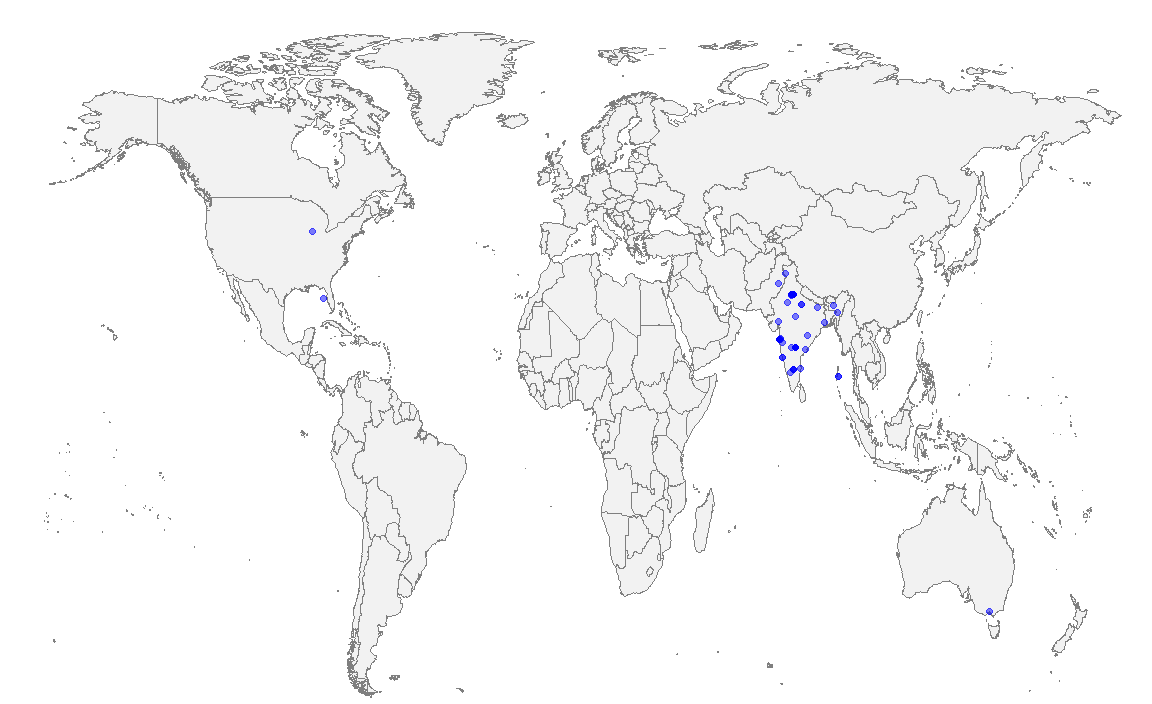


It is observed that, out of 1499 tweets the screen name who posted the maximum is “KVajpayee” with a frequency of “47” followed by a user with near high frequency of “”43”.

## **VISUALIZATION OF DATA**

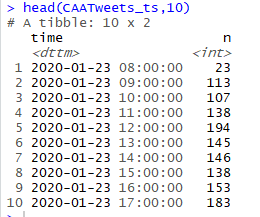
To get a better visual analysis of the tweets, an interactive map is created to show the location from where the tweets are being posted during the CAA event.

For this, the latitude and longitude information are extracted from the available geolocation information in the tweets, using which the below map is plotted.



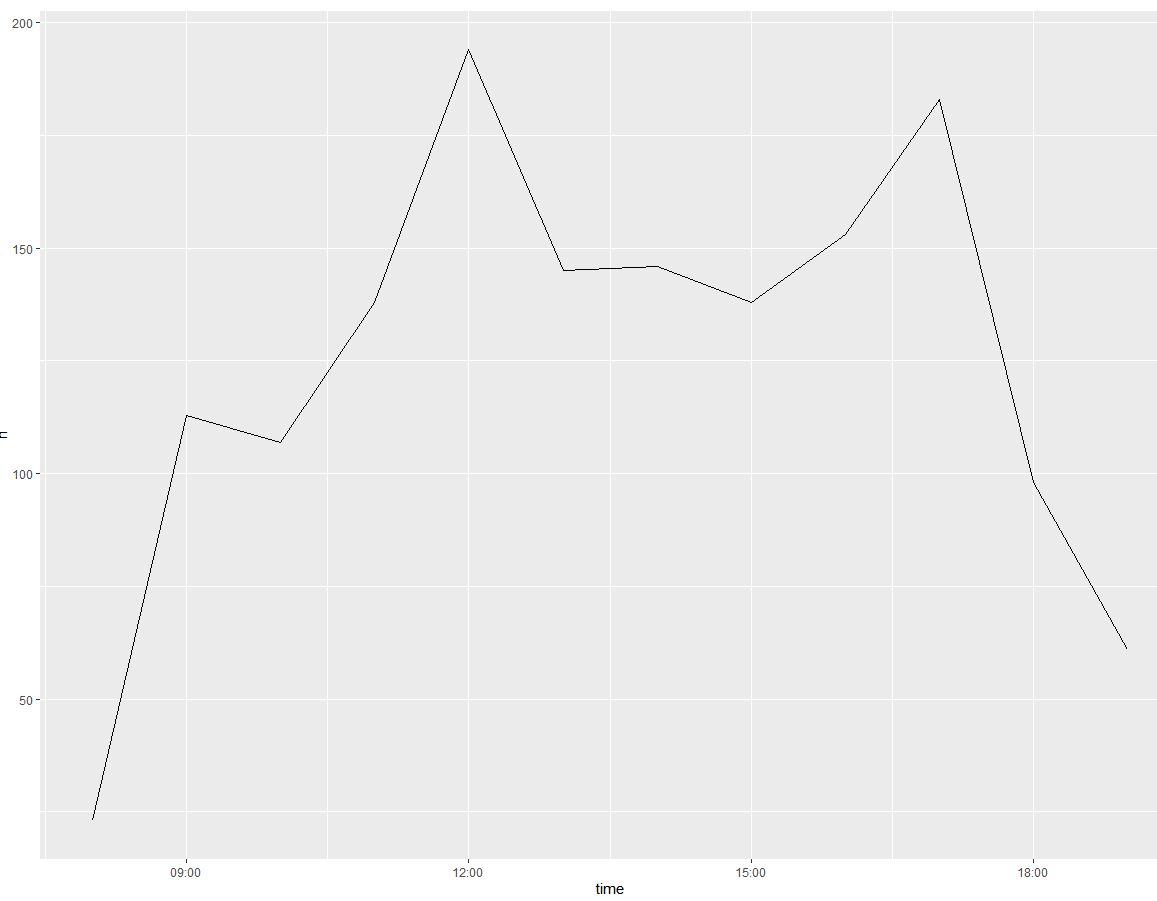
One can see that the maximum participation was from in and around **India**, and it is justifying as CAA in recent times is pertained to India.

Now, converting our tweets to a time-series like data object, the frequency of tweets is found out over a specific interval of time.



It provides us with the recent timeline, and this information will not be much in handy for further studies.

The peak time in hours of active participation is observed with a time series plot as below.



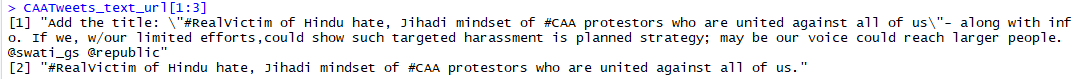
It is seen, around noon 12 and between 4 and 5 PM there is a high frequency of tweets being posted.

# **DATA CLEANING**

For this phase, the tweets column(text) are taken and saved into a dataframe to perform further cleaning.

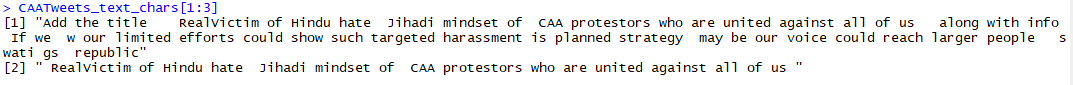
### **REMOVE URL’s**

First, the URL’s are removed from the tweets as they are not much needed for analysing the tweets alone. This is done with the help of “qdapRegex” package which consists of the rm\_twitter\_url() function.



### **REMOVE SPECIAL CHARACTERS, PUNCTUATIONS, NUMBERS**

Next, the special characters like “#,$,%”, punctuations(, . ?) and numbers are all removed at a go from the set of tweets as the text alone is the prior for our goal of mining.



Now, the dataframe is converted to a Corpus(collection of texts) in order to apply the “tm” package functions for further cleaning.

### **CONVERT TO LOWERCASE**

Then, entire corpus is converted to lowercase to avoid situations where for example, the model might treat a word which is in the beginning of a sentence with a capital letter different from the same word which appears later in the sentence but without any capital letter.



### **REMOVE DEFAULT STOPWORDS**

Further, the default Stopwords which are extremely common words like “and”, “or”, “not”, “in”, etc are removed. They are removed from the corpus as they provide no meaning or context for analysis.



### **REMOVE EXTRA WHITESPACES**

As observed, there are many extra whitespaces in between texts, which will in turn consume more space in the corpus. They are removed and made compact as below.



### **STEMMING**

As the final step of text pre-processing, stemming of the corpus is done. A word is replaced with its most basic conjugate form in stemming. For example, the stem of the word “typing” is “type”. We do this as we don’t want the words “type” and “typing” to convey different meanings to the algorithm.



In the above case, “protestors” and “united” are stemmed to “protestor” and “unit”.

## **ANALYZING TEXT FREQUENCY**

To understand our tweets even better and find the influence level, the frequency of each word is found out.

Top 50 most frequently occurring words and their frequencies are as listed below.

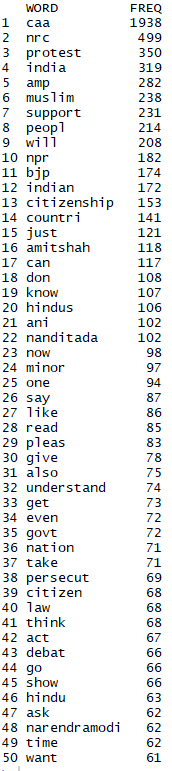
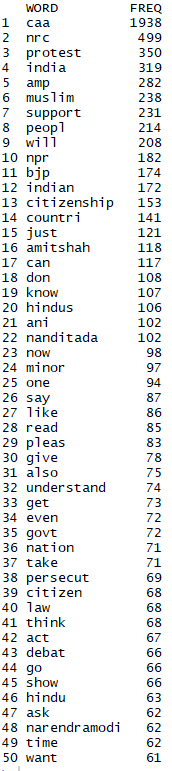
 



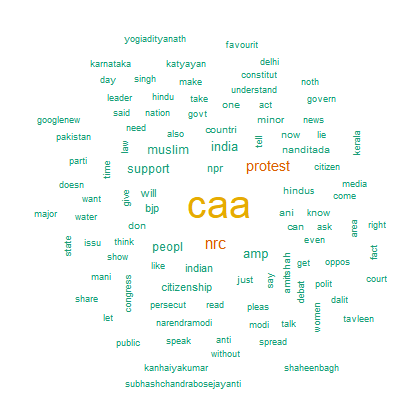
We understand that without any doubt CAA is the top occuring word followed by related words as NRC, protest and so on.

But also, few meaningless texts such as “s”,”t”,”u” seems to be unrelated and would accumulate more space as they account as high frequency terms.

So custom stop words are built and the unwanted words is omitted from the corpus. And the even compact corpus is shown below.

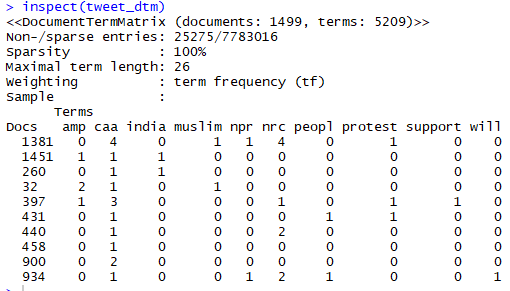
 

A word cloud is constructed to visualise the frequencies of all the above words. The larger the size, the higher the higher the frequency.



# **TEXT MINING**

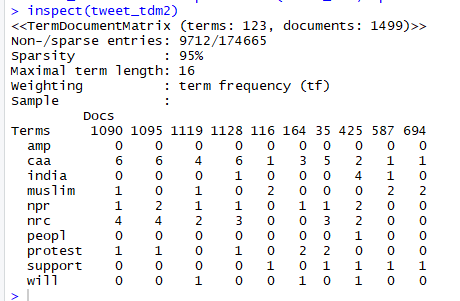
For further mining, we convert the Corpus to a Document Term Matrix and Term Document Matrix and make it a more compact and less dimension matrix.



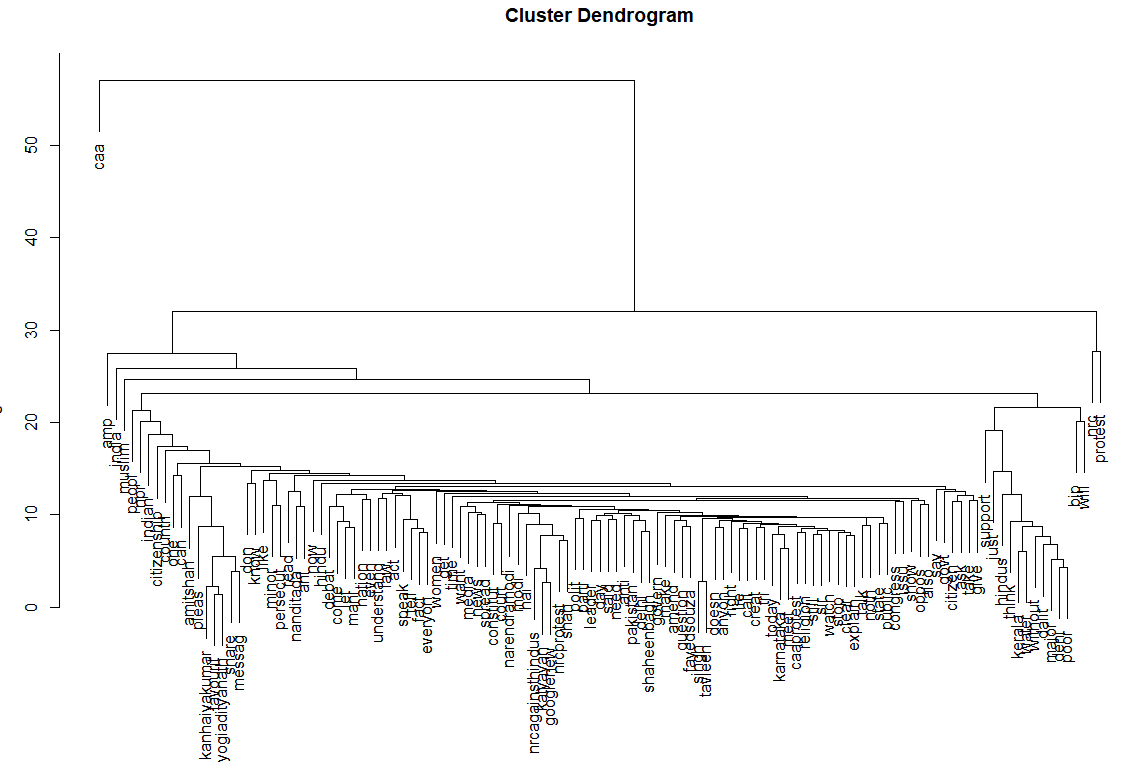
From the structure of the Document Term Matrix, it is observed that it consists of 1499 documents with 5209 terms.

The 0 values here which accounts to the sparsity factor are 25275 values.

Hence, we attempt to remove the sparsity and reduce the value, using the ***removeSparseTerms()*** function for the text data.



A **Dendrogram** is now plotted to understand the groups formed using our texts by hierarchical clustering.



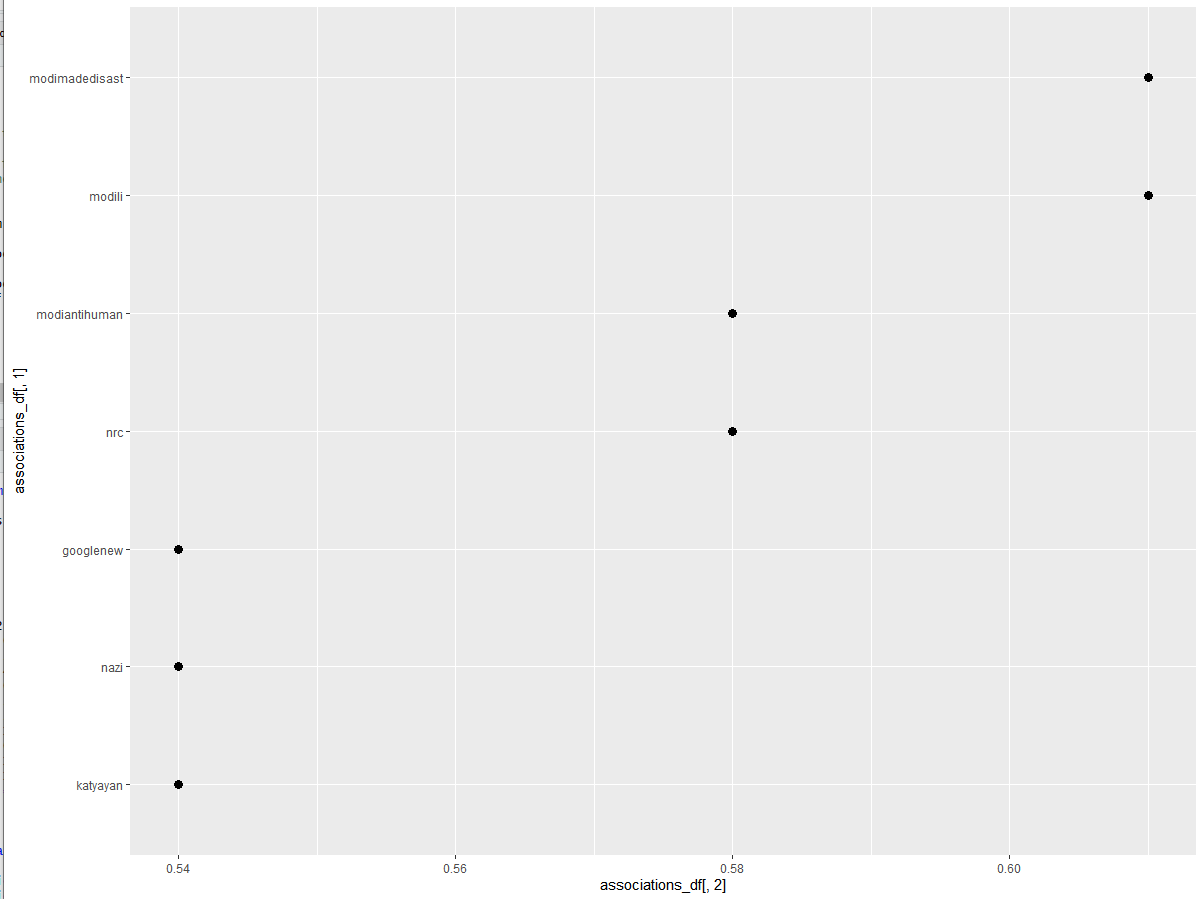
The hierarchy formed isn’t clearly visible, due to the higher number of terms (detailed view can be done in attached R code).

CAA being the root word leading to NRC, Protest and Amp and such clusters is observed.

## **WORD ASSOCIATION**

Here high frequency words are taken, to check the correlation using ***findAssocs().*** A condition of 50% and more association with the given word is being plotted.

So, first “CAA” is taken here and the below association plot is formed for it.

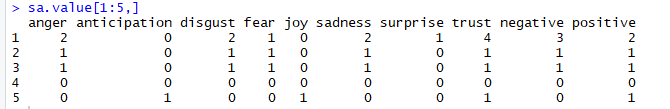


We can observe that the word has higher association with “**modili**” and “**modimadedisast**”. We come to few conclusions as to how influential are such associations found on the subject “CAA”.

# **SENTIMENT ANALYSIS**

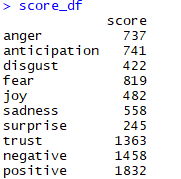
Now the sentiment is to be extracted from every row of terms / tweets so as to analyse how well and not well accepted was the Topic CAA among the public / screen names.

This is done with the help of “**syuzhet**” package which consists of ***iconv()*** to convert into characters recognizable to extract sentiment and ***get\_nrc\_sentiment()*** to find out the sentiment scores and the associated sentiment of the row.



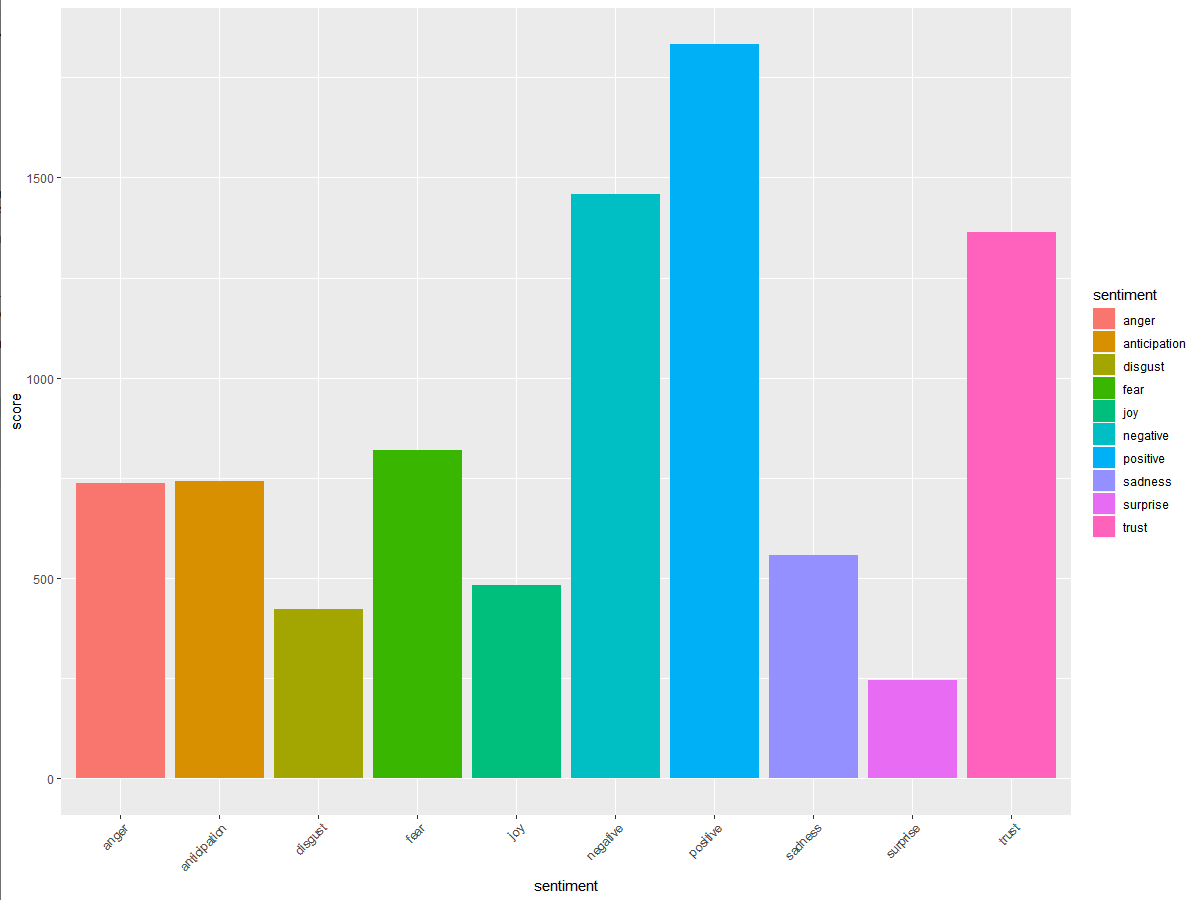
As shown, every sentiment is allotted a value respective to the tweet. For instance, the first tweet analysed was “RealVictim of Hindu hate mindset of CAA protestors who are united against all of us”. For the mentioned characters of 1st tweet, the emotions predicted were “Anger”, “Disgust”, “Fear”, “Sadness”, “Surprise”, “Trust”, “Negative”, “Positive” with varying values.

Now the sentiments extracted are collaborated, a collative sentiment on the whole is taken as shown below.



It is observed that completely Positive and Negative tweets put on are of count 1458 and 1832 respectively. Which accounts us to conclude that there is **more positive** participation in regards with CAA than negative.

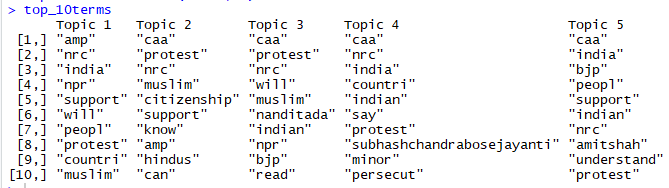
The sentiments and their value on our data are plotted for better visualization.



## **TOPIC MODELLING**

Now, similar topics are formed by grouping important/relevant words together across the documents. The different topics are extracted in which each consists of similar relevance and meaning.

Here 5 Topics are extracted with each 10 terms.



It is observed that, **Topic 4** is more pertained to a country perspective with the acts and personality. With detailed study, such topics can be used for understanding different groups of subject and its influence.