TEXT ANALYTICS ON 'SELF DRIVING CARS'

INTRODUCTION

Self-driving cars are a booming technology in today's world, they are presently the latest trend in the manufacturing industry that is going to revolutionize motoring. Its definition describes it as "A self-driving car (driverless car, autonomous car, robotic car) is a vehicle that is capable of sensing its environment and navigating without human input". It is said that this newest technology added in cars would help potentially save about 30,000 lives a year. Major players in self-driving car world are Google, Tesla, Uber, Apple, Audi, Intel and Baidu to name a few.

Why is it so interesting? It has been estimated that over 90% of car accidents on the road happen due to human error. If this human error can be replaced with advanced technology, many lives could be saved along with saving millions of dollars on accidents. These self-driving cars could also be used for picking up other passengers which would result in fewer cars on the street. They would also know better on hot to use fuel efficiently. These are some of the advantages to name a few.

Even after all the advantages listed out, customers still wonder on whether these cars would be 'too safe' to drive. Off late, Tesla has made its cars semi-autonomous along with Uber and Google bringing up their new self-driving car cabs. A lot is said and being said about these self-driving cars on social media in the form of comments, concerns, statements as well as suggestions. Some of these are positive, negative or even neutral.

Therefore, using text mining we seek to classify the opinions of the public on Twitter and understand what exactly the customers like or dislike, their current satisfaction and dissatisfaction about this futuristic 'driverless' technology. Sentiment analysis on various locations further performed can help us to gauge whether this technology is a success or a failure amongst the crowd at that location. We can also locate the terms that have been most frequently used with regards to the self-driving cars and check how they are affecting customer decisions.

DATA COLLECTION

The data collected was from Twitter by using TwitteR package. The package helped us extract tweets using rest API.

- *Duration of the data collected:* 1st Jan 2016 –Present
- *Total number of tweets extracted*: 6000
- *Keywords used to search tweets:* "self driving", "autonomous cars", "autonomous vehicle", "self driving cars"
- *Locations used:* California, New Jersey, Colorado, Delaware, New York, Arizona, Massachusetts

  Selected the major big cities where self-driving cars are being tested or have most talked about opinions on them

The extracted tweets were in csv format. The search results primarily consisted of attributes such as "text", "latitude "and "longitude".

PREPROCESSING TASKS

There were certain tasks to be performed to get a cleaner dataset for further classification.

Manual Annotation:

The tweets extracted did not have a sentiment attached to it. The extracted information just had only the tweet and location values. Each of the tweets were manually annotated in a spreadsheet format into "positive", "negative" and "neutral". These files were then ready to be pre-processed further.

Data processing tasks:

We considered only the text portion of the tweets from the .csv file. We use the "tm"(text mining) package in order to convert the text into corpus file and perform pre-processing tasks. Removed the spaces between the words and converted it into a Vector Source file. The vector source file was then converted into a Corpus object.

The various data processing steps performed on the corpus are:

- *Converted to Lower case:* Here the words in the corpus are converted into lower case by specifying attribute "tolower" inside the tm_map function
- *Removing Punctuation:* The text in the tweets might have various punctuation marks such as "?,:;!@()]\" to name a few. It is important to remove them, so we get a clean set of data that is easier to analyze with just the words. Here the punctuations are removed by specifying attribute "removePunctuation" inside the tm_map function
- *Removing Numbers:* There is a possibility of numbers being inside the text, by specifying the attribute "removeNumbers" inside the tm_map function, the numbers within the corpus will be removed.

- *Removing stopwords:* There are many stopwords that are in the text which would not give accurate results when analyzed as we need the important contextual words. The stopwords are removed by specifying the attribute "removeWords, stopwords("english")" within the tm_map function.
  The above attribute would remove the general stopwords in English, but we can also customize the stopwords.
- *Stemming words:* Stemming is the process of reducing derived words to their word stem, base or root form—generally a written word form. There are certain words in the tweets which have similar meaning and can be stemmed into one word. Used the attribute "stemDocument" within the tm_map function.

Converting Corpus to Document Term Matrix:

We used the function called as DocumentTermMatrix which converts the corpus into matrix form wherein each word is mapped to its document and calculated each word frequency in a decreasing order. Here we get list of most frequent words from the matrix in a decreasing order.

EMOTICON ANALYSIS

Emoticons are a new way to display user's sentiments on certain opinions to give added value to the statements. It was interesting to see what all emoticons were used and how can they play an important role in displaying sentiment towards self driving cars.

We used an emoticon csv file that has a pre-defined list of all emoticons, then in the code we pre processing the text content of the tweet for each of the location data. With the help of emoticon csv file, we mapped the emoticon and sentiment of location file onto a new file that had the extracted emoticons.

We were able to observe that certain emoticons were true to the sentiment of the statement made in the tweet and some were totally vague. We concluded that its not always reliable to base your sentiment analysis on emoticons, and it is necessary to take a look at the text content to derive the actual sentiment.

Please refer the appendix for the list of emoticons extracted.

WORDCLOUD VISUALIZATIONS

A wordcloud is a visualization of word frequency in a given text as a weighted list. The word frequencies are plotted as a cloud of words where the highest frequency word would be displayed in a larger font followed by decreasing size of words around it.

With the help of wordcloud visualization we were able to visualize the most frequent words within the tweets. Here we can see the trending words in the tweets that most people are talking about. Word cloud visualizations give an overall idea of the sentiments of what the audience thinks about self driving cars.

In this project the wordcloud displayed has frequent words such as "car", "drive", "self driving","california", "uber","help","baidu" and many others

Words such as california, uber and baidu represents the trending news on those words such as:

California: "Apple gets permit to test self-driving cars in California"

Uber: "Uber puts self-driving cars back on the road following crash"

Baidu: "Baidu, the "Chinese Google has its first fully self-driving car hit the road in China"

We used the package "wordcloud2" to perform the visualization. The wordcloud function takes in the words and frequencies as its attribute and displays the wordcloud

SENTIMENT ANALYSIS

We performed sentiment analysis of the Tweets using two different techniques:

1) Using Manually Annotated Tweets
2) Using an R package 'Syuzhet' that performs Sentiment Extraction and Analysis

1) MANUAL TEXT CLASSIFICATION

We manually annotated about 3841 tweets out into three categories: *Positive, Negative and Neutral*. Positive tweets were those that showed an optimistic or constructive emotion in their tweets, Negative tweets were those that showed undesirable or pessimistic emotion in their tweets and Neutral tweets were those that were basically facts or statements or those which had a balances positive and negative sentiment.

We used a classification-based approach where we used two different classification algorithms '*Support Vector Machine'* and '*Maximum Entropy'* Algorithms to train the classification model

and perform the predictions. We divided the dataset such that 70% of the data was considered as Training data and the remaining 30% was considered as testing data.

Using the Support Vector Machine algorithm, we were able to obtain accuracy of about *73.45%* which was quite reasonable enough.

Details on the performance of the SVM algorithm are as below:

CONFUSION MATRIX			
	NEGATIVE	**NEUTRAL**	**POSITIVE**
NEGATIVE	150	42	15
NEUTRAL	23	297	102
POSITIVE	29	153	560

STATISTICS BY CLASS			
	NEGATIVE	**NEUTRAL**	**POSITIVE**
SENSITIVITY	0.7426	0.6037	0.8272
SPECIFICITY	0.9512	0.8578	0.7378
POS PRED VALUE	0.7246	0.7038	0.7547
NEG PRED VALUE	0.9553	0.7945	0.8140
BALANCED ACCURACY	0.8469	0.7307	0.7825

Using the Maximum Entropy algorithm, we were able to obtain accuracy or about *69.58 %* which was little lower as compared to the performance of the SVM algorithm.

Further details on the performance of the Maximum Entropy algorithm are as below:

CONFUSION MATRIX			
	NEGATIVE	**NEUTRAL**	**POSITIVE**
NEGATIVE	153	39	15
NEUTRAL	25	303	94
POSITIVE	33	211	498

STATISTICS BY CLASS			
	NEGATIVE	**NEUTRAL**	**POSITIVE**
SENSITIVITY	0.7251	0.5479	0.8204
SPECIFICITY	0.9534	0.8545	0.6806
POS PRED VALUE	0.7391	0.7180	0.6712
NEG PRED VALUE	0.9502	0.7366	0.8267
BALANCED ACCURACY	0.8393	0.7012	0.7505

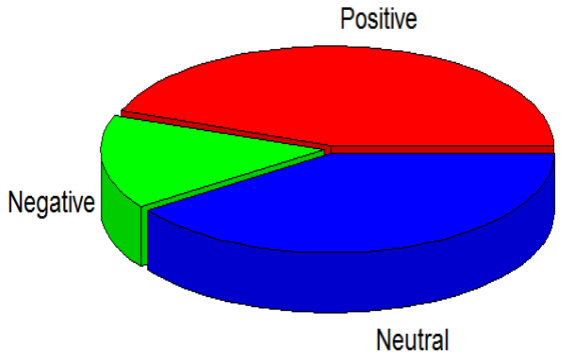
Some examples of the tweets classified using the Maximum Entropy Algorithm

"Great story and very cool product Looking forward to seeing this impacts AutonomousVehicles"	POSITIVE
"The iPhone of cars Apple enters the selfdriving car race via "	POSITIVE
"Who wants selfdrivingcars Millenials lead the way 40 are willing to use fully autonomousvehicles"	POSITIVE
"If Uber is depending on driverless taxis to become profitable it s doomed"	POSITIVE
"More millennial excited about flying cars than selfdriving cars"	POSITIVE

"Check out this site about driverless cars made by Brown students driverless selfdriving ai"	NEUTRAL
"From Tractor Trailers to Self Driving Machines"	NEUTRAL
"Overview of AutonomousVehicles competition between selfdriving"	NEUTRAL
"Navigation repo is 1st study that was done outside of selfdriving bandwagon Ford GM Waymo are real leaders"	NEUTRAL
"Varying by age group growing distrust in SelfDrivingCars has some data"	NEUTRAL

"That s horrible with tight streets as it is I don t trust driverless cars"	NEGATIVE
"I love driving I don t want driverless cars Ever"	NEGATIVE
"Serious privacy AND security risks lie in the path of connected driverless cars"	NEGATIVE
"Driving is purely the most boring thing to do ever Selfdriving cars get here soon enough car future"	NEGATIVE
"In 15 years millions of people will give up their cars for autonomous ride hailing"	NEGATIVE

Sentiment Distibution

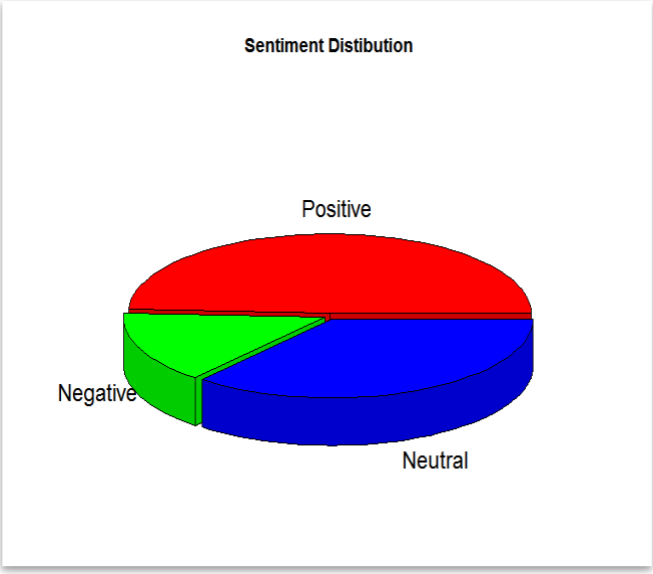


Some examples of the tweets classified using the Support Vector Machine Algorithm

"Apple getting into self driving game CA allows road testing autonomousvehicles"	POSITIVE
"Big Fuel Savings From SharedUse Of AutonomousVehicles"	POSITIVE
"Apple has a permit to test AutonomousVehicles in California"	POSITIVE
"selfdriving cars are a viable solution to the recent terror attacks using vehicles"	POSITIVE
"Driverless cars can revolutionize independence for the blind and disabled"	POSITIVE

"TECH NEWS How will maintenance change with the autonomous vehicle"	NEUTRAL
"The World s First Electric Driverless Roborace Racing Car"	NEUTRAL
"Driverless cars could either be scary or great for the environment"	NEUTRAL
"Self driving vehicles in the United Kingdom"	NEUTRAL
"How do you feel about self driving cars"	NEUTRAL

"Tesla s Autopilot self driving system slammed in lawsuit"	NEGATIVE
"Americans still wary of self driving cars an AAA study finds"	NEGATIVE
"When it comes to selfdriving young consumers more worried than excited"	NEGATIVE
"5 Ways Autonomous Vehicles Are Disrupting Healthcare"	NEGATIVE
"Most Americans Are Too Afraid to Ride in SelfDriving Cars via"	NEGATIVE

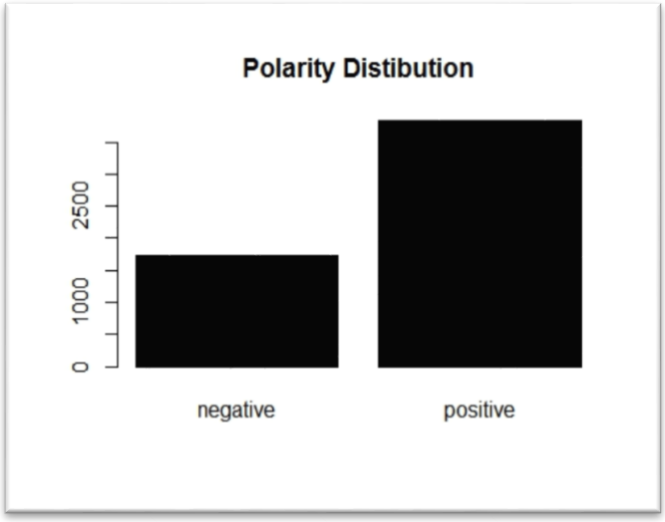


2) SYUZHET PACKAGE

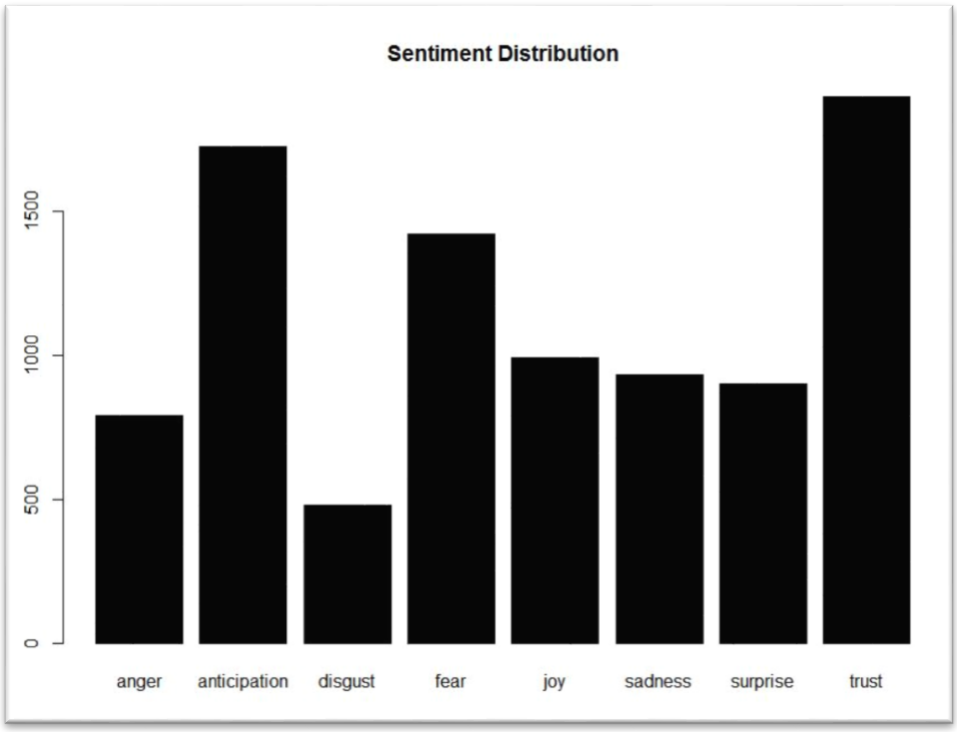
The syuzhet package includes four sentiment dictionaries and serves as a sentiment extraction tool developed in the NLP group at Stanford. It helps to extract sentiments and sentiment-derived plot arcs from text using the sentiment dictionaries such as 'Bing','AFinn','NRC' and 'Syuzhet(default)'. "Syuzhet" was developed in the Nebraska Literary Lab, "afinn" was developed by Finn A. Nielsen, "bing" was developed by Minqing Hu and Bing Liu and "nrc" was developed by Mohammad, Saif M. and Turney, Peter D.

Initially we used the NRC dictionary in order to extract the sentiments as well as the polarities. The polarities extracted are categorized as Negative and Positive and the sentiments were further categorized into Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust.

NRC Sentiment Distribution Figure:

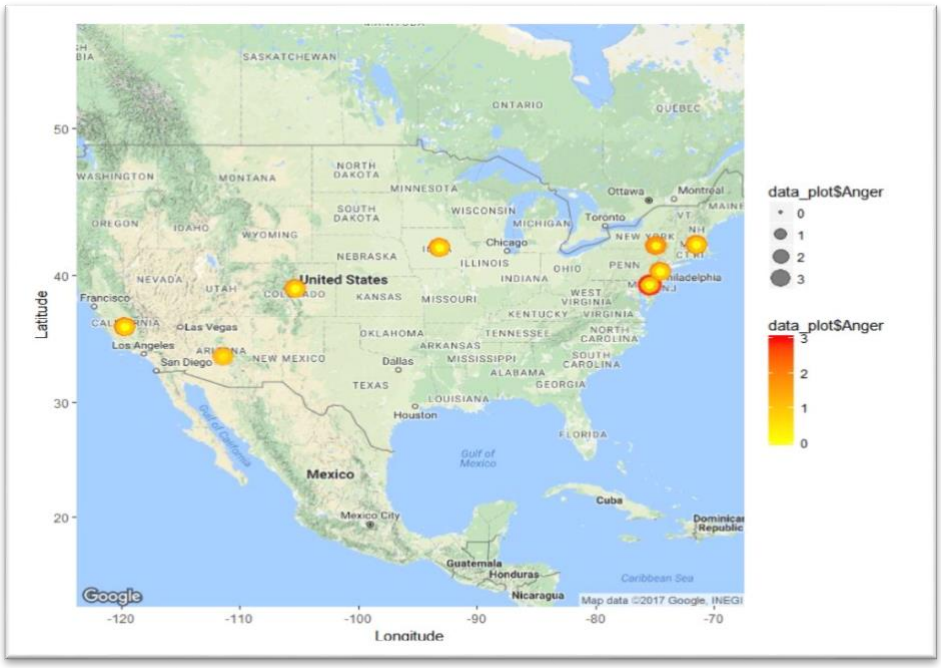


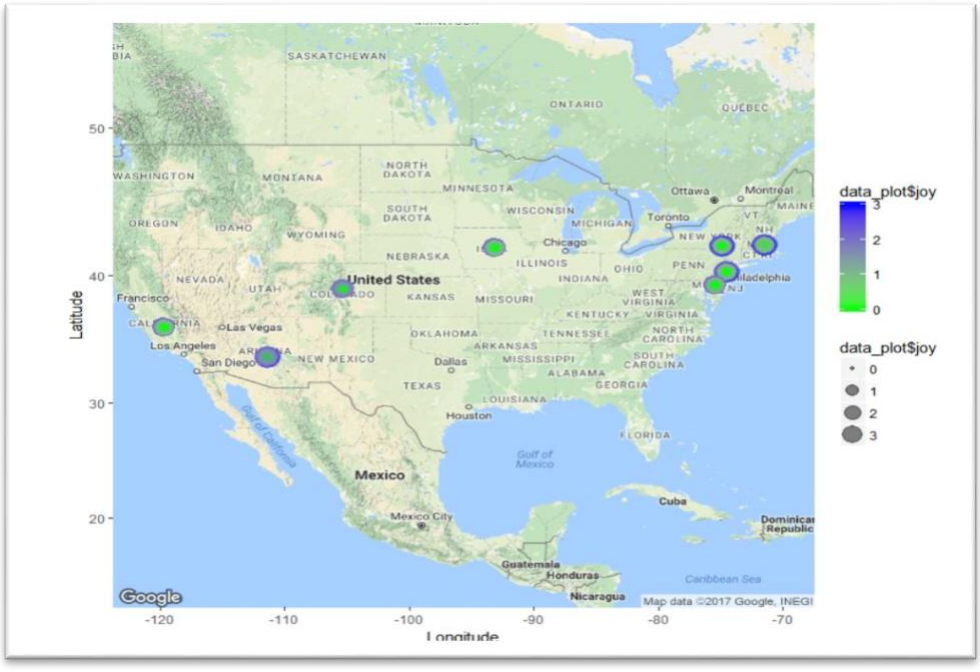
NRC Polarity Distribution Figure.:

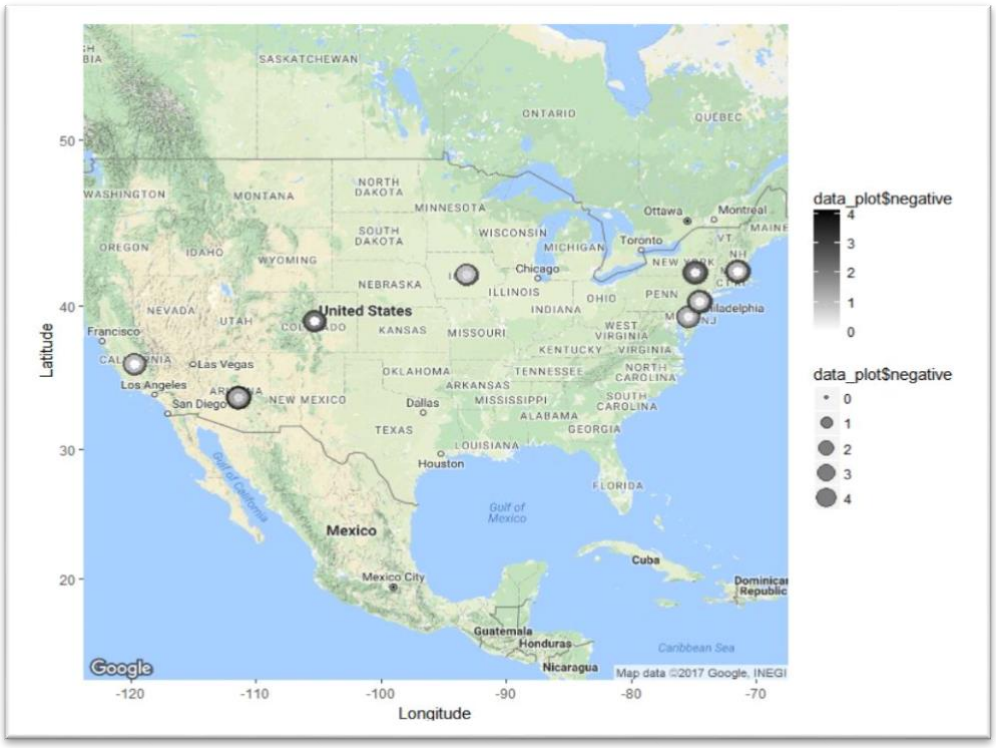


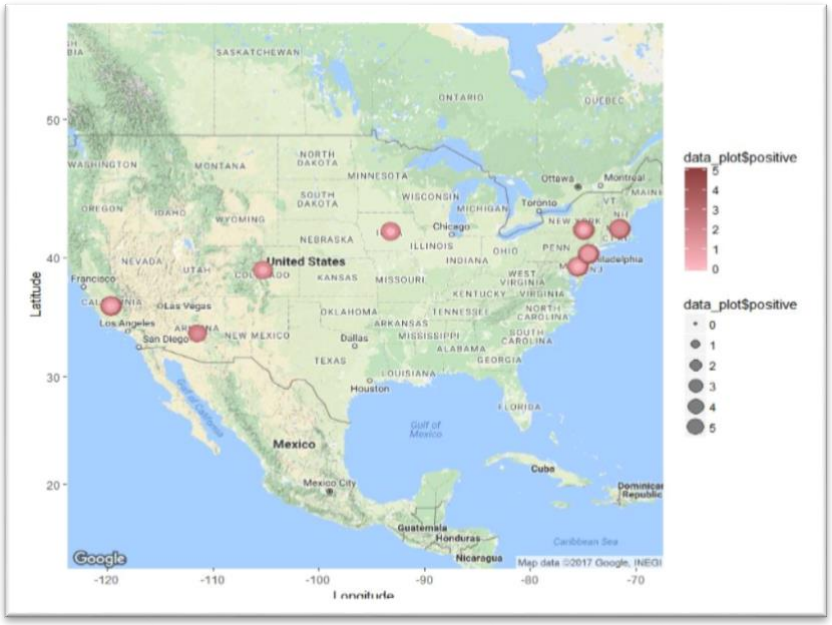
Further location wise visualizations were also plotted for each of the sentiments, such as anger, joy, disgust, trust etc. It was done using the ggmap package in R.

Location-Wise plots for Anger, Joy, Negativity and Positivity Sentiments:


data_plot$Anger
0
1
2
3
data_plot$Anger
3
2
1
0
Latitude
Longitude
50
40
30
20
-120
-110
-100
-90
-80
-70
United States
Mexico
Cuba
Chicago
Toronto
Ottawa
Montreal
Philadelphia
Las Vegas
Los Angeles
San Diego
Dallas
Houston
Mexico City
Guatemala
Honduras
Nicaragua
Gulf of Mexico
Caribbean Sea
Google
Map data ©2017 Google, INEGI


data_plot$joy
3
2
1
0
data_plot$joy
0
1
2
3
Latitude
Longitude
50
40
30
20
-120
-110
-100
-90
-80
-70
United States
Mexico
Cuba
Chicago
Toronto
Ottawa
Montreal
Philadelphia
Las Vegas
Los Angeles
San Diego
Dallas
Houston
Mexico City
Guatemala
Honduras
Nicaragua
Gulf of Mexico
Caribbean Sea
Google
Map data ©2017 Google, INEGI



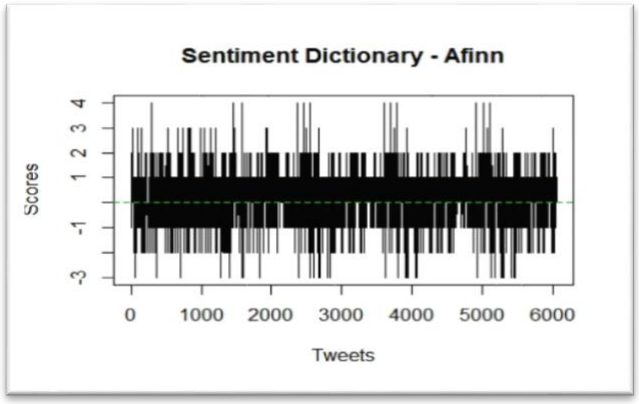


All of the above analysis helped us answer the research question defined in our initial project report viz. geographically, what are the variations in opinions and sentiments in order to determine its prevalence and impact globally in our case across certain states within USA?

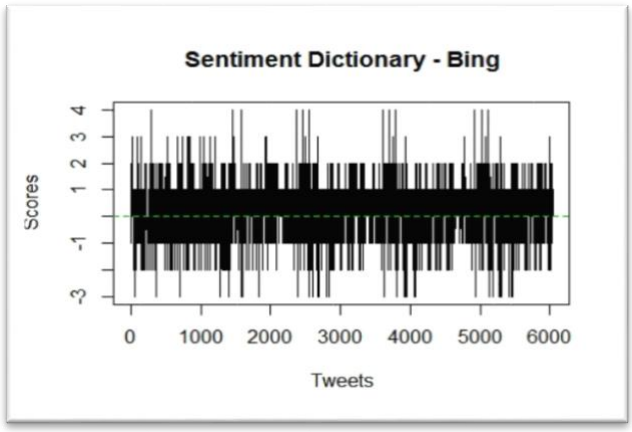
Post this; we also decided to explore the other sentiment dictionaries that were available within the Syuzhet package.

We used each of the sentiment dictionaries that would help extract the sentiment based on different sentiment scored defined within the methods. We then plotted graphs that would help us visualize the sentiments across the entire set of tweets.

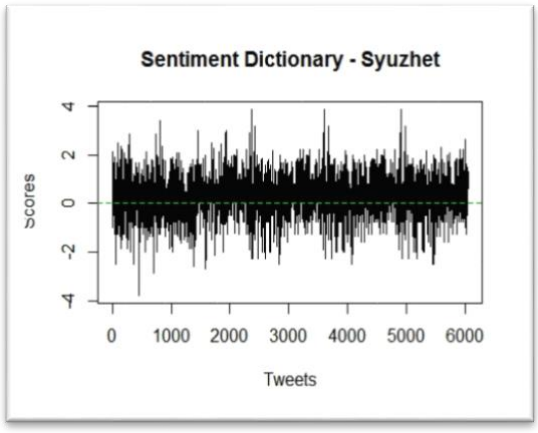
Sentiment Dictionary-Afinn graph:



Sentiment Dictionary-Bing graph:

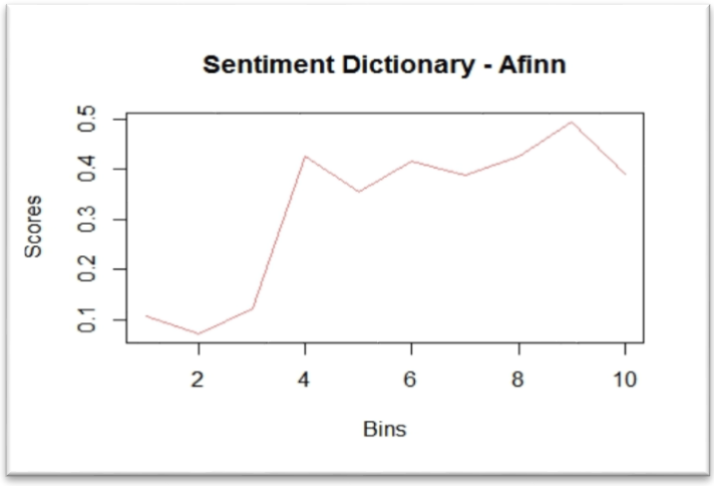


Sentiment Dictionary-Syuzhet graph:

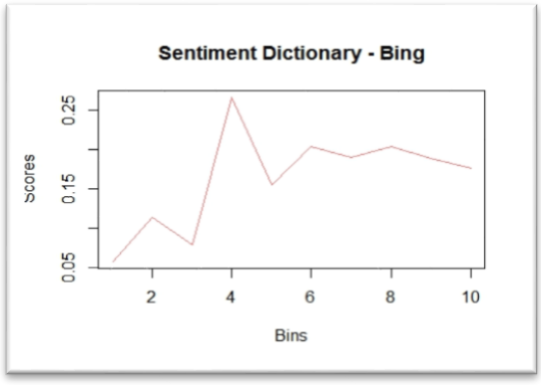


The graphs obtained using the procedure defined above seems to be noisy and is not very appealing to the eye. We decided to enhance the visualization a little more and hence we went ahead by dividing the data into chunks or bins. We decided to go ahead with 10 bins, and then visualize the overall sentiment with respect to those bins.

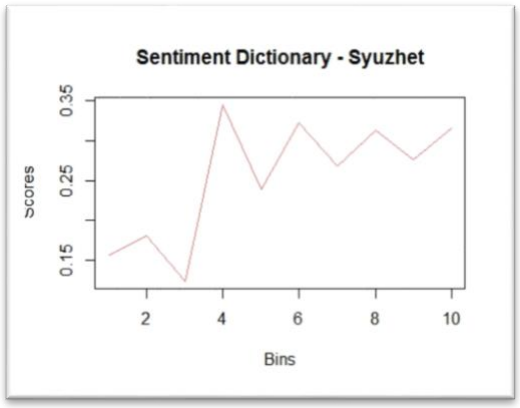
Sentiment Dictionary-Afinn graph using Bins:



Sentiment Dictionary-Bing graph using Bins:



Sentiment Dictionary-Syuzhet graph using Bins:



CONCLUSION

In conclusion to this project, we were able to answer the questions presented in the initial report such as:

- How data from Twitter can be analyzed in order to understand the generic opinions – positive and negative of 'Self-driving cars' based on public opinions?
- Geographically, what are the variations in opinions, trending news and sentiments in order to determine its prevalence and impact globally?
- How to identify the various dimensions of interest with regards to the opinions in terms of keywords and use visualization to do so?

The future work for this project would involve in exploring and analyzing a greater number of tweets from various locations within USA and understand the sentiment overall. Dwell into more visualization techniques to view the analysis done on the dataset.

Topic modelling could be another area to explore wherein we can obtain tweets based on popular users and understand their opinions on self-driving based on the topic generated.