Time based analysis of Tweet sentiment
Case Study : Gun-Reform

Bhushan Jagtap, Kartik Joshi, Vishal Kudale

***Abstract*—Allow user to provide a hashtag as input and gather all the data related to given hashtag (such as event specific hashtag E.g. #GunReform). Perform sentiment analysis on extracted tweets and plot that on time series format. Along with Time series data plotting, plot geo-location data, Devices used for tweeting, and information like count of unique users, tweets likes, re-tweets etc. As part of dynamic implementation user can select a date and according to selected date by user, plot before and after sentiment analysis for a given event. The main purpose of this implementation is to determine how people have reacted to specific event, at what extent they used tweets to express their thoughts and how the response has been formed over the time. Also meta information analysis help us to have a good visualisation on many technical aspects.**

I. Introduction

FOR this implementation we kept our scope to only one event and which is how people react to gun violence and gun reform act. This can be used for any other events. How US people reacting on twitter with respect to gun violence. In this project we have try to capture reaction and trend of people on twitter with respect of gun reforms after incidents. Our assumption is that in recent years many people have moved towards having strict gun rules so that it can prevent such shootout incidents. We will try to capture that movement in form of positive tweets from the people who wants change. Also there is one hypothesis that says people are active to such talk or changes for few days or month after incident and after some time they move on to some other topic. In this project we will capture that in the form of time series plotting of trend of given topic over given time window. In recent years people are using twitter to express their views and feelings related to big topics like terrorist attack, gun violence, passing a law in congress etc. The analysis only covers how people are reacting to such gun violence incidents. The main reason to perform this analysis is to see how people react to such event on and how it is changing over time. We have implemented and plotted different graphs to understand each aspect of data science. List of plots and their importance listed below:

1) Time Series Analysis for Sentiment
   This graph shows how peoples views and sentiments changes with respect to time. A plot which shows how people used twitter to express their negative and positive views related to a specific topic.
2) Before-After analysis
   To see what was the response of people before and after certain date. To capture a change in reaction for a specific issue with respect to a date. So we can see how a specific event has changed peoples reaction and their response on twitter.
3) Date-wise plotting of twitter attributes
   Get day by day analysis of attributes like, Total number of tweets, Retweets, Likes and how many unique user have tweeted that day. This will help us to understand how much people are interested in given issue. Also how many new people have joined to express their feelings on twitter.
4) User Location Plotting
   To see whether people from specific region is tweeting about given issue or people across united states are actively participating in given discussion. This will help us to understand awareness of people with respect to given problem and also to track location wise activities of people on twitter.
5) Source Device Plotting
   Plot the source of device used to tweet. This will give us an technical aspect of analysis. What medium people use to reach out to outer world to express their concerns.
6) Total Positive-Negative Sentiment
   This is a plot that consider sentiments for all the tweets related to a topic so far to visualize whether the overall response is positive or negative.

All this plots are helpful to visualize and make conclusion whether people have a positive or negative response with respect to a specific event or concern. Also time series analysis will help to get a detailed overview on trend, connectivity, reach and involvement with respect to time.

II. Overall Architecture

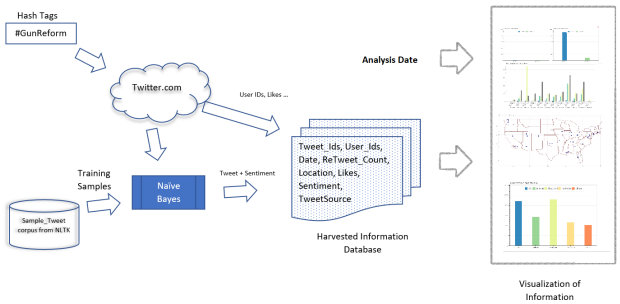

Fig. 1. Application Architecture

The above diagram depicts overall architecture our project. Following are steps our project will perform.

1) User Enters Hashtag
   We have provided a text box where user can input all the hashtags on which he wants to perform time-based analysis.
2) Gathering information from twitter and perform sentiment analysis
   Once we have all the hashtags user want to search, We will query to twitter using tweepy API to get all tweets related to entered hashtags.
3) Pre-process data and sentiment analysis
   Once we have all the information related to hash tags we will extract following fields from it:

   `Tweet_Ids`, `User_Ids`, Date, `ReTweet_Count`, Location, Likes, Tweet Source, Actual tweets

   Then we will perform sentiment analysis on tweets and we will store all information along with tweet sentiment in database for further analysis and visualization. (Detailed information on sentiment analysis is given below)
4) Acceptance of analysis date from user
   We have provided a drop-down list from which user can select a date on which user wants to divide gathered information. This division of tweets will be used to analyze the effect of selected date on tweets and user sentiment.
5) Visualize results
   Once we have all the processed information we can visualize the results based on date entered by user using graphs such as time series, geo-plot, sentiment analysis bar-graph etc. (Please check visualization section for detailed information).

III. Data Pre-processing

Tweepy API search method returns collection of tweet objects in the json format.
Extracted fields which are of our interest but some of them were not in the required state. So we did following pre-processing on the data before storing it into CSV file:

- Twitter has saved geo-location in bounding box of coordinates which encloses the place. It is series of longitude & latitude points, defining a box which will contain the place entity this bounding box related to. So, we have taken the average of longitude & latitude separately to get a single coordinate point.
- To get the source of the tweet, we have parsed the source attribute from the json & compared parsed data with specific words like iPhone,Android,Web-client to decide the source of the tweet & accordingly saved the data as a likert scale.
- As retweet count attribute present in the tweet object is same for actual tweet & retweets of that tweet. So, if we consider retweet count of each tweet object then we will have wrong retweet count. E.g.: Consider a tweet X & Y people has retweeted this tweet. In this case, retweet count for X & all Y will have value as Y.Here, actual retweet count is Y but if we will get (Y+1)*Y. So, to get the actual retweet count, we first checked each object whether is it a tweet or retweet? If it is tweet then only we considered the retweet count in our data.

IV. Sentiment Analysis

We are performing sentiment analysis on tweets gathered for some time say 10 years to analyze the effect of various events. We are using Nave Bayes classification technique to classify each tweet into positive or negative tweet. Along with being simple Naive Bayes is also useful for large datasets.
Using Nave Bayes Theorem, we can calculate posterior probabilities P(C—X), from class prior probability P(C) and likelihood of each attribute as show in following equation:

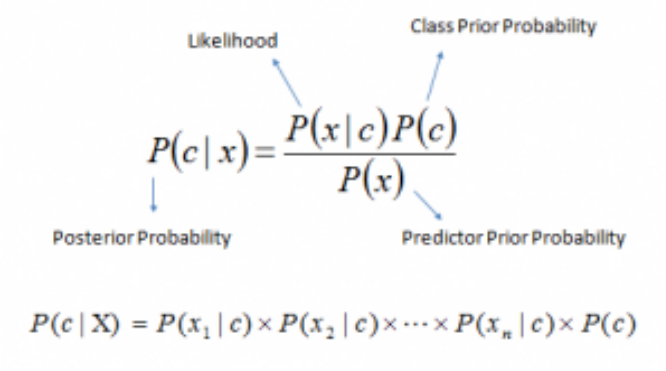

Fig. 2. Bayes theorem

Above,

- P(c|x) is the posterior probability of target class c given predictor attribute x.
- P(c) is the prior probability of target class.
- P(x|c) is the likelihood which is the probability of predictor given target class.
- P(x) is the prior probability of predictor.

Following are key attributes of our sentiment analysis implementation:

1) Data set
   We are using **sample_tweets** corpus provided by NLTK package for training Nave Bayes Classifier. This corpus contains total 10000 sample tweets in raw format along with positive and negative sentiment labels. Our code divides these tweets into two parts in 80:20 ratio for training and testing respectively.
2) Accuracy of Nave Bayes
   Our current implementation provides 96% accuracy on sample_tweet corpus test set.
3) Preprocessing of raw tweets
   The data provided in sample_tweet corpus and data we get from twitter is in raw format and need to be processed before feeding it to classifier. In our current

implementation we are taking care of following cases:

a) Emoticons(E.g. :) :)) etc.)
b) StopWords (E.g. A, An, The etc.)
c) Case Sensitivity
d) URLs, Hashtags, Usernames
e) White Spaces and Repeating Characters

4) Feature selection and training
We are using most common words found in training set as feature for our classifier. Along with most common words we are converting URLs, Hashtags and emoticons into keywords and feeding it to Naive Bayes classifier. We believe that emoticons are great indicator of users emotion and should be used for sentiment analysis.
5) Train once approach
As everyone knows, training classifier is time consuming task. Hence our code comes with pre trained classifier trained on sample_tweets corpus saved in pickle object. Also, we allow user to retrain the classifier. Once retraining is done we save trained classifier in pickle object in classifier folder. So whenever user inputs new hashtags for analysis our code will pick up stored classifier saving us from wasting time on training phase.

V. Visualization

As mentioned above in abstract, this implementation allows to perform analysis on any event, public concern, or any ongoing discussion on social platform. User can provide respective tweet hashtag in input and the follow steps to get the detailed results.

Fig. 3. Input Box for user

You can use above mentioned text input to specify the hashtag for which you want to perform the analysis. Once you provide that and run remaining cell in jupyter notebook, you will get the all the charts at the end. Here due to twitter api restriction we can only extract data for last 10 days. Because of that all the results mentioned below are generated on data extracted for #GunReform for last 10 days.
List of graphs generated as part of this analysis

1) Time Series Analysis for Sentiment plot
This graph shows how peoples views and sentiments changes with respect to time.
As you can see graph in Fig.4, contains date wise analysis of positive and negative tweets. This time series data allows us to understand the trend of tweets with respect of time and their sentiments. By visualizing time series data we can say whether people have same sentiment throughout the entire time or it has changes, whether it become more positive or negative, Is there any

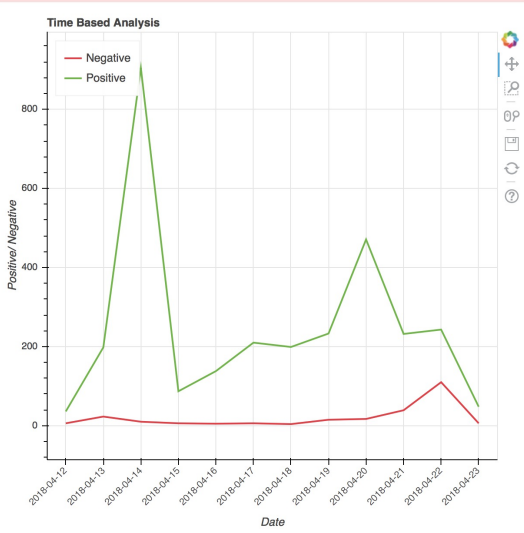

Fig. 4. Time Series Analysis

spikes in daily results due to some event. As you can see in result we have data from April 12th to April 23rd 2018. From this plot we can say that people have positive views when it comes to strict gun reform. More number of positive tweets then compare to negative tweets.

2) Before-After analysis plot
To see what was the response of people before and after certain date.

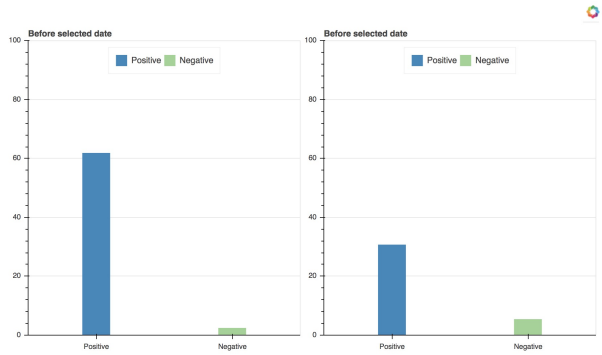

Fig. 5. Date based Before-After analysis

Fig. 6. Date Selection drop-down

One of our hypothesis that certain event can have impact on views of people. Ex. Shootout at a school might lead increase in tweets related to strict gun reform act with positive sentiments. On that hypothesis we build this graph such that if we want to see how people were reacting on twitter before and after certain date for a given issue. By selecting date from drop down two

graphs will be generated dynamically. Both containing total number of positive and negative tweets.

3) Date-wise plotting of twitter attributes
   Get day by day analysis of attributes like, Total number of tweets, Retweets, Likes and how many unique user have tweeted that day.

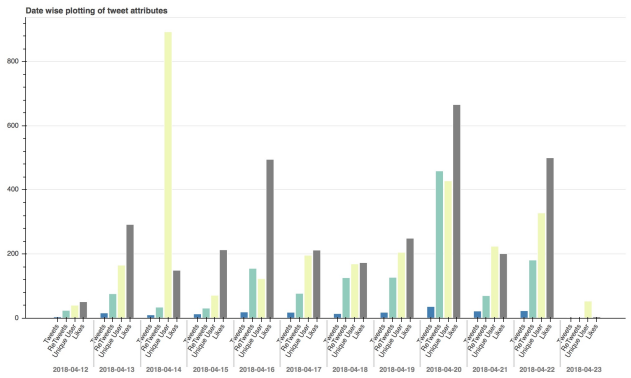

Fig. 7. Date based comparison of twitter attributes

This plot is used to get general information like how many tweets, and retweets related to specific topic on given day along with likes for tweets and retweets. Also this plot contains list of unique user. This will help to analyze the trend how new people reacting to specific problem. This statistical information is helpful to get a clear picture on how many people are involved, concerned and actively taking participation on given discussion.
This will help us to understand how much people are interested in given issue. Also how many new people have joined to express their feelings on twitter.

4) User Location Plot
   To see whether people from specific region is tweeting about given issue or people across united states are actively participating in given discussion.

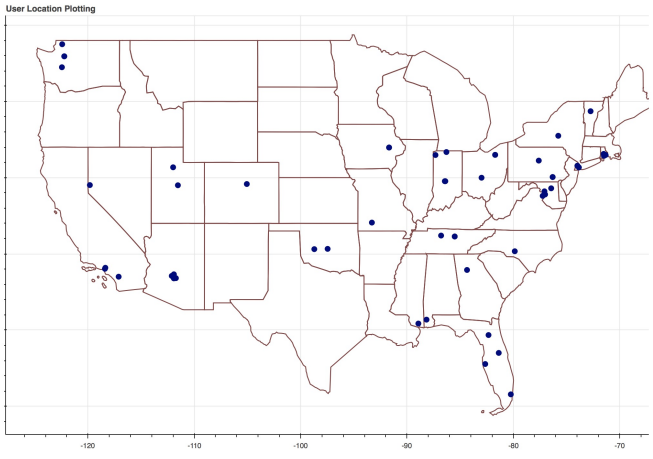

Fig. 8. User's Location Mapping

To access the location on twitter user has to enable location sharing. Once its enabled twitter stores the location where given tweet was tweeted.This will help us to understand awareness of people with respect to given problem also to track location wise activity of people on twitter. As mentioned above twitter api will only generate data for last 15 days also less people allow access to their locations, because of that we only managed to plot few of the tweet origin. But with enterprise twitter api with same code we can get more tweets and get more data points on given plot.

5) Source Device Plot
   Plot the source of device used to tweet.

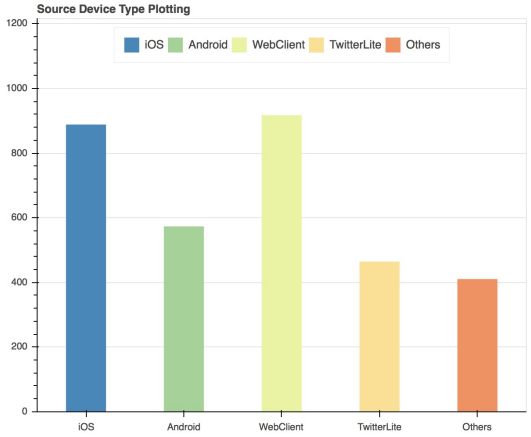

Fig. 9. Devices used to tweet

This will give us an technical aspect of analysis. What medium people use to reach out to outer world to express their concerns. To see whether people use phones to tweet or they use webportal to express their concerns. Android, IOS app and Twitter Lite can be accessed via phone. From graph we can say that more number of people use mobile device to tweet compare to personal computer as they are handy now a days.

6) Total Positive-Negative Sentiment Plot
   This is a plot that consider sentiments for all the tweets related to a topic so far to visualize whether the overall response is positive or negative.
   Fig. 10. gives details on overall response of people for given time frame by ploing total positive and negative tweets.From this we can say that there are more number of positive tweets compare to negative tweets.

VI. Technical Details

- Packages used
  1) Bokeh
     We have used Bokeh 0.12.7 version. We have used bokeh to take user input from user. We have used it to display different plots like time-series plot, bar charts, geo-location plot.
  2) Tweepy Library
     We have used Tweepy library to extract the tweet related data from Twitter.

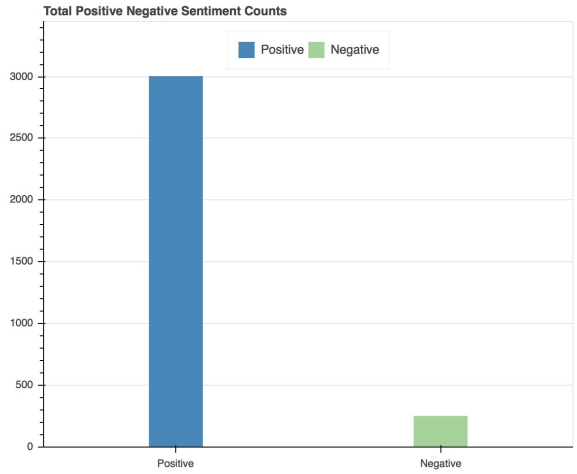

Fig. 10. Overall Sentiment Analysis

3) NLTK Library
This library is used to train our classifier based on twitter_sample data. It is also used to perform sentiment analysis over the tweets extracted from twitter based on users input.

- Dependencies
Our application is dependent on following:
1) Bokeh
2) Numpy
3) Blaze
4) Dask
5) Pandas
6) JsonPickle
7) Tweepy
8) NLTK

Run following commands to install above dependencies on Anaconda Prompt:
1) python -m bokeh info
conda uninstall bokeh
conda install bokeh=0.12.7
2) conda install NumPy
3) conda install blaze
4) conda install dask
5) conda install pandas
6) pip install jsonpickle
7) pip install tweepy
8) pip install -U nltk

VII. Issues faced

Twitter api(Tweepy) has access to data just for last 10 days.So we have changed our scope of the project from overall analysis to analysis on available last 10 days data.

VIII. Conclusion

We are able to perform a sentiment analysis on extracted twitter data based on the submitted query(hashtag) by user. Also We have represented all sentiment results in the form of time-series graph & bar charts. With the help of these output, one can definitely decide responses of people from all over the world to given query i.e. one can easily say whether peoples response is positive by just by observing the output plots. From our analysis, there is a positive response from people for gun-reform act law. People are using mobile devices than web-client to tweet their thoughts. It is not the case that people are active only from the place where gun-shooting incidents are happened but people are active all over the United States. Also, it is not the case that only some people are just tweeting/retweeting about gun-reform act. There are some new people who also tweets on it.

References

[1] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[2] http://docs.tweepy.org/en/v3.5.0/
[3] http://bokeh.pydata.org/en/0.9.3/docs/user_guide/charts.html
[4] https://bokeh.pydata.org/en/latest/