Title

DATA VISULAIZATION AND SENTIMENT ANALYSIS ON TWEETS ABOUT KASHMIR ISSUE

Targeted 62 politician from 10 different countries, local tweets from 8 different countries, local tweets from 3 larger cities of Pakistan and local tweets form 3 larger cities of India

Abstract

Today is the digital age of Social Networking. All the social sites are at peek these days. Some of the top social sites like LinkedIn, Facebook, Twitter, Instagram are stepping forward day by day and have billions of users Facebook is the most used social platform with 2.45 billion users, Instagram has about 1 billion users, while Twitter has about 330 million users and LinkedIn has  310 million users. There are many more social websites with millions of active users daily, generating thousands of terabytes of data daily. Social websites encourage users to express and exchange their views, suggestions and opinions publicly. The data created also contains these opinions and reviews concerned to a wide range of topics, they may be about a product, or a service, or even more important topics related to government and other social issues.

Twitter has steadily attracted users to convey their opinions and perspective about any issue, brand, company or any other topic of interest. For this reason, twitter is used as an informative source by many organizations, institutions, companies and even governments. Governments can use sentiment analysis to observe the effect of tweets on people. Consumers to research products or services before making a purchase. Marketers can use this to research public opinion of their company and products, or to analyze customer satisfaction, public view of government policies etc. Organizations can also use this to gather critical feedback about problems in newly released products.

Twitter has become a media of delivering messages and issuing statements for political figures and other dignitaries and public figures. The messages, better known as tweets, are sometimes positive and negative. Thus leaving information of people thoughts and opinions.

This paper highlights Kashmir issue and analysis on it by tweets of 62 politicians form 10 different countries, local tweets from 8 different countries and local tweets from 6 larger cities of India and Pakistan.

This research is achieved by a new area of research called data analysis and sentiment analysis. We are making sure to provide a comparative visualization and sentiment analysis of tweets of politician, countries and cities based on number of tweets, fav count, retweets, time period, polarity and subjectivity.

Keywords

Twitter, Sentiment Analysis, Data Analysis, Tweepy, Pandas, Getoldtweets3,TextBlob, opinion mining, power BI, aspect based mining, extraction, cleaning, polarity, Excel, nltk,

Data Visualization & Sentiment Analysis

Sentiment Analysis is the automated process of analyzing text data and sorting it into sentiments positive, negative or neutral. Performing Sentiment Analysis on data from Twitter using machine learning can help companies understand how people are talking about their thoughts. We are required to formulate that what kind of features will decide for the sentiment it embodies.

The basic steps for performing sentiment analysis includes data collection which can perform by crawling the tweets against target hash tags, pre-processing of data(cleaning), analyzing tweets for sentiment and performing classification either using simple computation or else machine learning approaches. After careful evaluation, we can visualize the results on tweets.

Performing sentiment analysis is challenging on Twitter data. The extraction and analysis of huge unstructured internet content is beyond the human power and time. Problems arise also because of limited tweet size, which causes users to use slangs, abbreviations and emoticons. These require different processing than other words.

Our target is to analyze the tweets about Kashmir issue. We gather data of most followed politicians of the world. The target is to analyze that how many countries actually talk on this topic and how many countries are in favor or against of this issue. In recent years, the Kashmir issue arises between India and Pakistan. many politicians have started to use twitter to express their feelings about this issue. When any politician tweets about something their tweet is highlighted that what are their thought on this topic. Either they are in favor, against or neutral. and this is the main  purpose of our research paper.

Statement

* Our first target was to extract tweets of 62 politician and perform data visualization and sentiment analysis ,for this we gathered data for top 10 politician from Pakistan, top 10 from India, top 5 form Bangladesh ,top 6 from Canada, top 7 from Iran , top 6 from United kingdom, top 6 from USA, top 6 from Saudi Arabia , top 6 from Turkey and we got only 1 profile for Palestine
* Our second target was to perform data visualization and sentiment analysis on tweets generated from 8 different countries, for this we gathered data of *Bangladesh, Canada, Germany, India, Pakistan, Turkey, United Kingdom and USA*
* Our third target was to perform data visualization and sentiment analysis on tweets generated from 6 top largest cities ,3 from Pakistan and 3 from India. For this we gathered data of *Karachi, Lahore, Islamabad, Mumbai, Bangalore and Delhi.*

In recent years people have started to use twitter to express their feelings and share opinion. the main purpose of research was to read behavior and thoughts of people in last 6 months about Kashmir issue.

.

Methodology

In order to perform sentiment analysis, we make use of different Python libraries that are considered giants in their area. The most renowned among their families of functionalities are Text Blob and Tweepy

1. Extraction Of Tweets - (Tweepy,Pandas and GetOldTweets3)

When we start our project the first step is to collect the data from twitter. We had collected our data from twitter API. We first went with Standard Twitter Search API which only returned data for past 7 days using tweepy libaray. Tweepy is basically a python library for access the twitter API. This library allow the user to gather the information after authentication. This library is basically used for download twitter status i.e. tweets in real time. To extract data for past 6 months then we then moved to GetOldTweet3 that is A Python 3 library and a corresponding command line utility for accessing old tweets. (<https://pypi.org/project/GetOldTweets3/>).

1. We first installed project to extract required data through <https://pypi.org/project/GetOldTweets3/> using following commands.

 pip install GetOldTweets3  
or pip install -e git+https://github.com/Mottl/GetOldTweets3#egg=GetOldTweets3

1. Then we used few commands to extract particular data

Eg:

GetOldTweets3 --username "imrankhan" –-queryseacrh ”kashmir ” --since 2019-08-01 --until 2020-01-31 -- --emoji unicode

Pandas is very famous Python library which is very easy to use data structure and gives high performance and the main thing is analysis our data very good, it is very fast and efficient for manipulation with integrated the indexing. It gives facility to read and write between the use of memory and use different formats like csv, sql database and text files.

1. Translation Tweets - (Google API)

As we selected tweets from various countries, we had to convert all tweets to English before going for sentiment analysis, The main language used to tweet by many people is in English. But tweets that were from turkey,Saudi Arabia,Bangldesh and japan were not in English .this was a problem as our sentiment analyzer is based on English. For this, we use Googles Translator that simply translates our text in Urdu, Turkish, Japanese to English that was then passed on to the Textblob for analysis

1. Storing Tweets In GITHUB

The tweets extracted were in csv format, which after cleaning was stored in local PC and then shared on GitHub so that later researchers can take benefit <https://github.com/iqraakhtar7/Data-Analysis-on-kashmir-Issue>.

1. Data Sentiment Analysis using TextBlob

TextBlob is a python library and offers a simple API to access its methods and perform basic NLP tasks.

A good thing about TextBlob is that they are just like python strings. So, you can transform and play with it same like we did in python.

Steps to go through

* 1. Install text blob in your system using the following command

pip install -U textblob

* 1. Download the text corpora, a large and structured set of electronically stored and processed texts using the following command

python -m textblob.download\_corpora

* 1. Perform NLTK tasks using text blob. For all nltk tasks, perform these three basic steps.
     1. Import text blob
     2. Create a textblob object and pass a string with it.
     3. Call functions of textblob in order to do a specific task.

Now lets specify funtion for each nltk task

* Tokenization use tokenize()
* Noun Phrase extraction use noun\_phrase()
* POS tagging use

for words, tag in blob.tags:

print (words, tag)

* Lemmatize use lemmatize()
* Creating N grams, N grams are helpful while working with combination of words. An n-gram contains n words, and treat them as single word. Use the function ngram(parameter) and pass the value of number of words needed to be combined to extract the value of a sentence. For example

Ngram(2) ; it groups very two words and refered as bigram

Using TextBlob we finally assign polarity to tweets. It is one of the challenging tasks to assign the polarity to a personal opinion, when you are working with machine.We work on two main aspects of a sentence, discussed as follows

**Polarity:** It is a float value which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement.

**Subjectivity:** Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1]. Where 0 means fact and 1 means personal opinion

1. Visual Representation of Analysis Using Excel, R And Power BI

1. In **Excel**, charts are used to make a graphical representation of any set of data. A chart is a visual representation of the data, in which the data is represented by symbols such as bars in a Bar Chart or lines in a Line Chart. Excel provides you with many chart types and you can choose one that suits your data or you can use the Excel Recommended Charts option to view charts customized to your data and select one of those.

We use sunburst, bar plot 2D, cluster bar plot 3D for our visualization.

2**. R** is a language and environment for statistical computing and graphics. R is also extremely flexible and easy to use when it comes to creating visualizations. One of its capabilities is to produce good quality plots with minimum codes.

We use some functions for visualization which are

* **plot():**

The plot() function is a kind of a generic function for plotting of R objects.

* **bar plot():**

In a bar plot, data is represented in the form of rectangular bars and the length of the bar is proportional to the value of the variable or column in the dataset. Both horizontal, as well as a vertical bar chart, can be generated by tweaking the horiz parameter.

* **Histogram**

A histogram is quite similar to a bar chart except that it groups values into continuous ranges. A histogram represents the frequencies of values of a variable bucketed into ranges.

3. **Power BI** is a cloud-based business analytics service from Microsoft that enables anyone to visualize and analyze data, with better speed and efficiency. It is a powerful as well as a flexible tool for connecting with and analyzing a wide variety of data. Power BI’s ease of use comes from the fact that it has a drag and drop interface. This feature helps to perform tasks like sorting, comparing and analyzing, very easily and fast. Power BI is also compatible with multiple sources, including Excel, SQL Server, and cloud-based data repositories which makes it an excellent choice for Data Scientists.

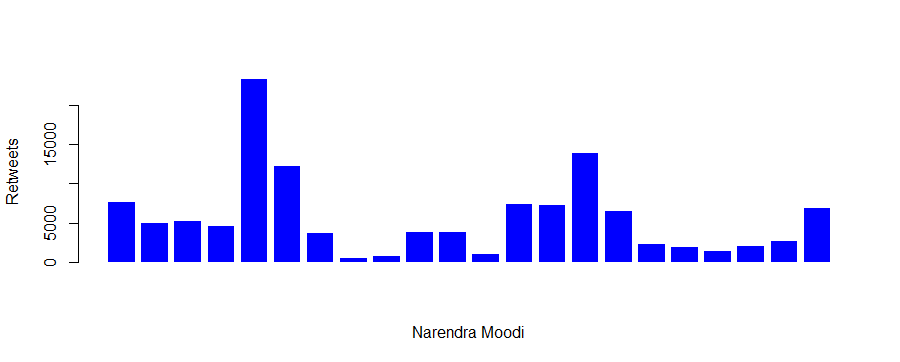
Power BI gives the ability to analyze and explore data on-premise as well as in the cloud. Power BI provides the ability to collaborate and share customized dashboards and interactive reports across colleagues and organizations, easily and securely.

Analysis

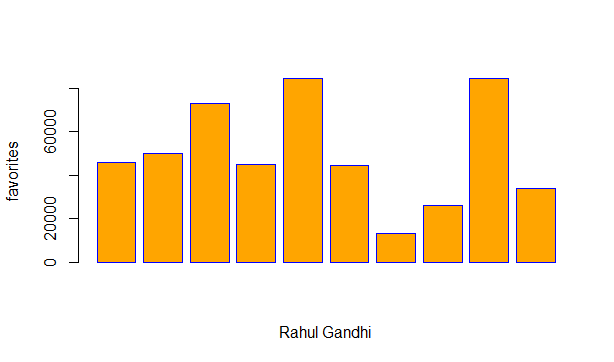
For our paper, we have analyzed the tweets of sixty two politicians from ten different countries and compare number of all tweets with number of tweets about Kashmir

1. plotted a graph for favourite count of tweet containing word kashmir
2. plotted a graph for retweet count of tweet containing word kashmir
3. We compared no of tweets with respect to time period(month)
4. Polarity and subjectivity of each tweet
5. Value graph that shows no of positive, no of negative and neutral tweets.

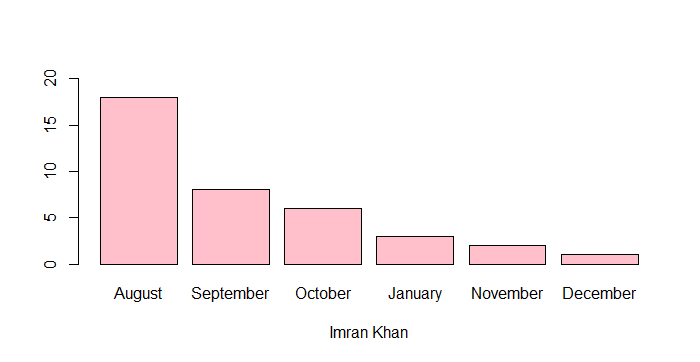
**Retweet**

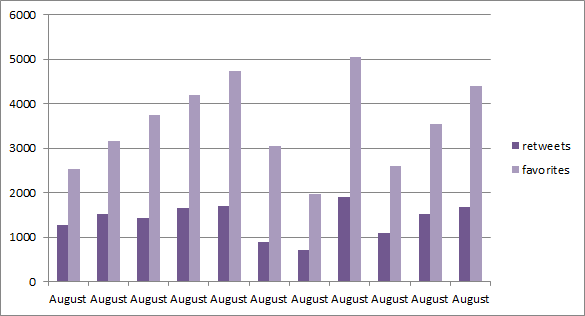


**Fav count:**



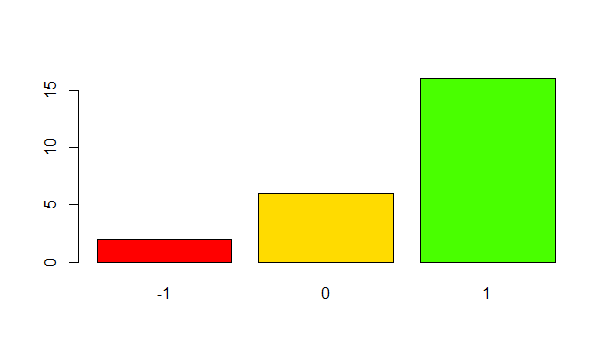
**Time(per month):**



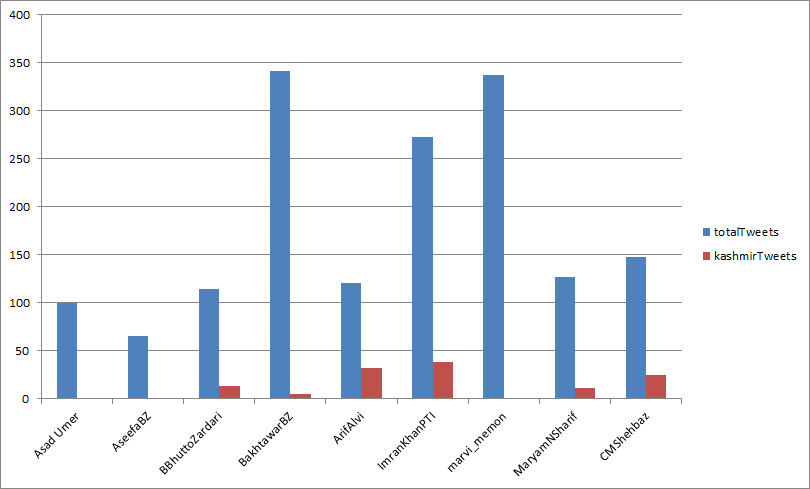
**Rtweet,favcount and time combined**:  


**Polarity:**

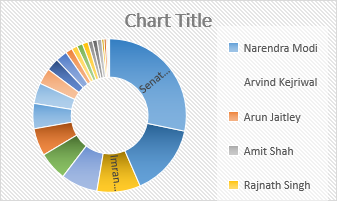
**Value graph(summarization of total number of positive negative and neutral tweets,-1 shows negative tweets ,1 shows positive and 0 shows neutral ):**



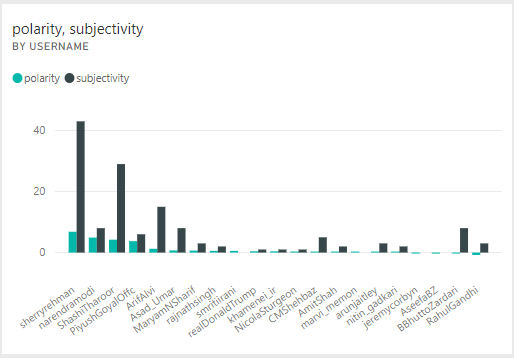
**No of all tweets compared with no of tweet a about Kashmir (country wise for politicans):**



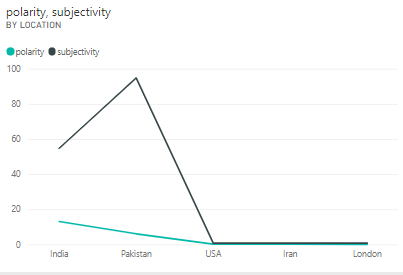
**No of tweets compared of all 62 politican:**



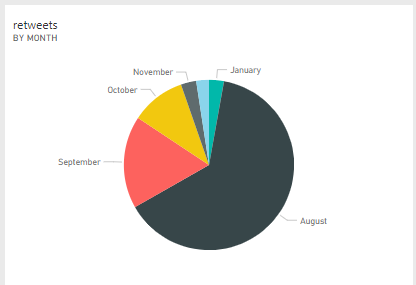
**Polarity and subjectivity of all 62 politicans:**

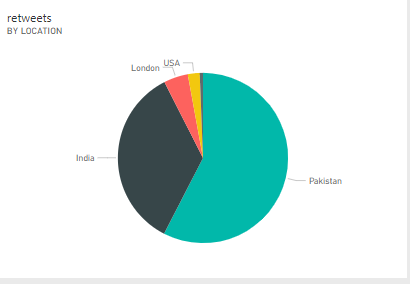


**Relationship between positivity and subjectivity (62 politicians):**



**No of retweets on the basis of month and loactions (62 politicians):**

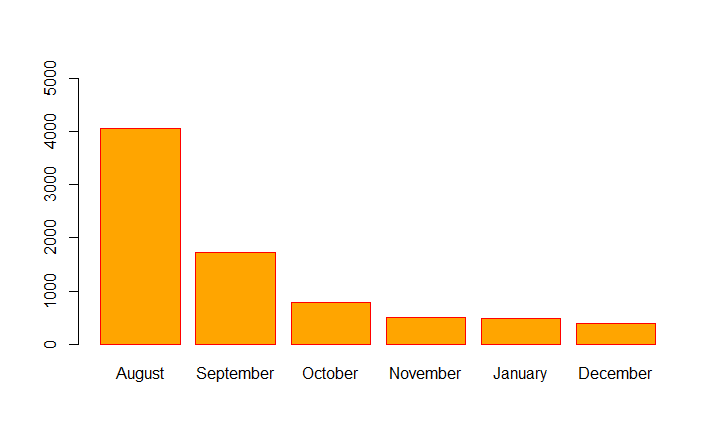


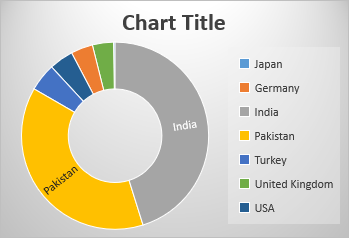


We gathered data of multiple countries (eight countries) from twitter to get the tweets about Kashmir for the time period of 6 months i-e august 2019 to January 2020

1. Plot a graph for number of tweets about Kashmir with respect to time period(per month) and country
2. Plot a graph for number of tweets about Kashmir with respect to time period(per month) (all countres combined)

**Uk**

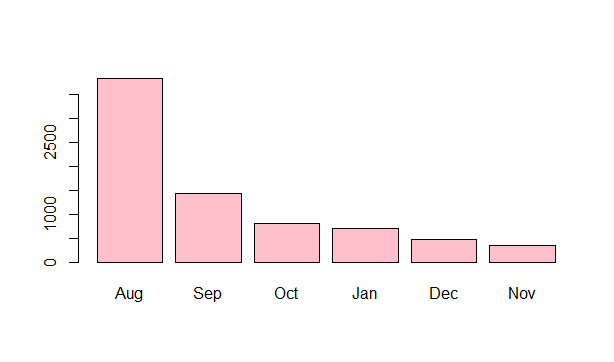




Third module was to gather data of three larger cities of Pakistan from twitter to get the tweets about Kashmir for the time period of 6 months i-e august 2019 to January 2020

1. Plot a graph between number of tweets about Kashmir from particular city and time period (per month)
2. Plot a graph between number of tweets about Kashmir from all 3 cities of on country and time period (per month)
3. Plot a graph for all 6 cities combined
4. Plot a graph to show ratio between number of tweets and population of particular city

**Mumbai**



**Combined cities within country**

**6 larger cities of india and pakistan**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Cities* | Total No.Of Tweets | Population | ratio | %Age of population who wrote about kashmir |
| *Karachi* | 3628 | 16,094,000 | 0.000225 | 0.022543 |
| *Islamabd* | 2869 | 1,129,000 | 0.002541 | 0.254119 |
| *Lahore* | 2874 | 12,642,000 | 0.000227 | 0.022734 |
| *Banglore* | 3504 | 12,327,000 | 0.000284 | 0.028425 |
| *Dehli* | 12736 | 30,255,200 | 0.000421 | 0.042095 |
| *Mumbai* | 7617 | 20,000,000.00 | 0.000381 | 0.038085 |

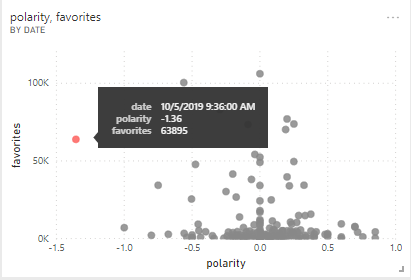
Here is a table that shows cities with their population, number of tweets, their ration and percentage, below is the graph that illustrates this statement.

Fourth module was to move data of sixty two politicians on Power BI dashboards. You can access the power BI dashboards using the following this [link](https://app.powerbi.com/groups/me/dashboards/e90881d0-6620-44b9-818f-e341c9a60d91?ctid=34104e4e-dc67-44c0-8e48-6eaa66937ed0)

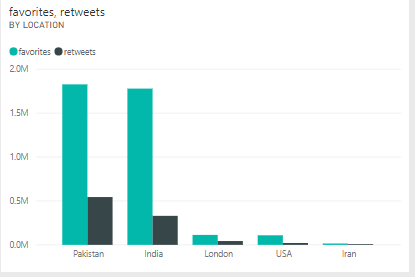
The dashboard is enabled for both Mobile and desktop views.

Most meaningful graphs are shown here

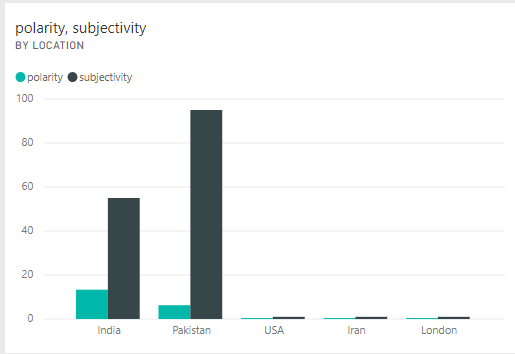
**When outlier has more number of favourites and a negative polarity**



**When Pakistan ha more retweets and favourites**



**Relationships between subjectivity and polarity**



Challenges Faced

We faced quite a few challenges in our research, regarding the extraction, cleaning and sentiments.

While extracting tweets we found that The Standard Twitter Search API can only search/fetch tweets published in the past 7 days and then discovered GOT package, which is useful in getting older data.

Using GetOldTweets library, and extracting tweets on the bases of countries ,we faced an issue that was must use of –within parameter so that we were than able to get realistic data, otherwise we were getting very less amount of data .we used parameter –within”25km” that covers radius of 25 km outside that country and city.

While using different python libraries we discovered that polarity is assigned to a specific word not the whole context, for opinion based mining, n-grams are recommended but they also works in the same ways. Our results are based on polarity of single word aggregated for each sentence. We use Power BI to analyze the output and conclude final statement

Future Work

First of all, the problems we faced were an obstacle in our research. Even though our work was automated, but we had to solve some problems manually and had to clean data and assign polarity to some of the tweets manually. In the future, when we would have rather larger amount of data, we would need to have better programs that would correctly analyze context and we mainly focus on methodologies that assign polarity with better accuracy.

Our research can be used to provide basic future works, which may include different opinion analysis, aspect-based or context based opinion mining.Reserachers can collaborate on writing algorithms on such data sets where negative words such as killing, blood shed are used in positive statements and how to assign the accurate polarity to such data sets.

Our next target is to use maps, such as Kohenn Maps, ANN, CNN and RNN and using new technologies such as Pytorch to analyze the data with best possible results.

Conclusion

So we conclude that now everyone can understand the power of social media so we should utilize it for better purposes.

We can conclude that our analysis is based on results obtained through algorithms and automated procedures without the interference of our personal opinion and we can say that according to number of favorites and retweets, we can fairly say that people of Kashmir deserves the right to live their lives according to their wills and killing and torture should be stopped to preserve peace.

References

* <https://github.com/Jefferson-Henrique/GetOldTweets-python>
* <https://pypi.org/project/GetOldTweets3/>
* <https://developer.twitter.com/en/docs/tweets/search/overview>
* <https://artelliq.com/blog/how-to-use-got3-to-extract-old-tweets-in-python/>
* <http://blog.chapagain.com.np/python-get-twitter-tweets-using-get-old-tweets-package/>
* <https://www.datamentor.io/r-programming/bar-plot/>
* <https://www.statmethods.net/graphs/bar.html>
* <https://developer.twitter.com/en/docs/tweets/search/faq>
* https://www.kdnuggets.com/2018/07/comparison-top-6-python-nlp-libraries.html
* https://elitedatascience.com/python-nlp-libraries
* https://towardsdatascience.com/self-organizing-maps-ff5853a118d4
* https://github.com/karanluthra/twitter-sentiment-training/blob/master/install.py
* https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed
* https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed
* https://pypi.org/project/flasgger/
* https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/
* https://github.com/vivekkalyanarangan30/Text-Clustering-API/tree/master/python
* https://dzone.com/articles/opinion-mining-python-implementation
* https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/
* https://github.com/abhinavralhan/kohonen-maps
* https://github.com/abhinavralhan/kohonen-maps#kohonen-maps
* http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/
* https://github.com/udacity/deep-learning-v2-pytorch/blob/master/sentiment-rnn/Sentiment\_RNN\_Exercise.ipynb
* <https://github.com/udacity/deep-learning-v2-pytorch>
* http://textblob.readthedocs.io/en/dev/