LIVE PLOTTING AND SENTIMENT ANALYSIS OF TWEETS ON A WORLD MAP

Aasawaree Deshmukh, Mohit Goyal, Rishabh Agrawal, Nishant Priyam

asdeshmu@syr.edu, mogoyal@syr.edu, ragrawal@syr.edu, npriyam@syr.edu

Abstract—**Twitter network is currently overwhelmed by massive amount of tweets generated by its users. To effectively organize and search tweets, users have to depend on appropriate hashtags inserted into tweets. We are performing historical and live analysis regarding the most talked hashtag or trending topic by searching the hashtag or word as an user input within a tweet and then doing a sentiment analysis of those tweets. Later, this data is plotted on the world map representing the sentiments on different continents with different colors providing an overview regarding the emotions about that topic. This will help various organizations or individual highlighting those regions (for eg. negative sentiment) which need improvement.**

I. INTRODUCTION

SOCIAL media are web-based communication tools that enable people to interact with each other by both sharing and consuming information [4], [6]. The popular social media platforms are obvious to most people - Facebook has an audience of some 2000 million, LinkedIn and Twitter both exceed 300 million, and Pinterest and Snapchat have close to 200 million each [7]. The reach of these platforms makes it clear that social media is now ubiquitous. All of these sites allow users to create and share content with an audience you haven't met in real life and connect with them.

Our problem statement is to give the user an holistic overview for the sentiment about a given trending topic and plotting the same on a world map. We analyzed various social media and networking sites which would help us achieve our goal. Facebook is a popular free social networking website that allows registered users to create profiles, upload photos and video, send messages and keep in touch with friends, family and colleagues. Twitter is a free microblogging service that allows registered members to broadcast short posts called tweets. LinkedIn is a social networking site designed specifically for the business community. The goal of the site is to allow registered members to establish and document networks of people they know and trust professionally. We explored few other social media and networking platforms as well but evaluated that Twitter would server as the best data source for our project.

II. OVERVIEW

A. Data Source

Twitter is a social networking and microblogging service, enabling registered users to read and post short messages, so-called tweets. Twitter messages are limited to 140 characters and users are also able to upload photos or short videos.

Tweets are posted to a publicly available profile or can be sent as direct messages to other users. Twitter is one of the most popular social networks worldwide. Part of the appeal is also the ability of users to follow any other user with a public profile, enabling users to interact with celebrities who regularly post on the social media site.

Due to the restriction on the number of characters in a tweet, people convey their thoughts or sentiments regarding a particular topic in precise and concise format. This aids us in collecting the data regarding a trending topic and in performing the sentiment analysis for the same. Due to all the above reasons, we choose Twitter as the data source for our project.

B. Twitter Module

Twitter's micro-blogging developer platform offers several tools and APIs. The Twitter REST API methods allow developers to access core Twitter data [1]. The Search API methods give developers methods to interact with Twitter Search and trends data. The API presently supports the following data formats: XML, JSON, RSS etc. It provides three tiers for searching Tweets - Standard, Premium and Enterprise. We decided to use the Standard search in comparison with other tiers as it is free of cost. It gives historical tweets up to a timeframe of 7 days.

We will be using a Python library called Python Twitter Tools to connect to Twitter API and for downloading the data from Twitter. The Twitter API requires OAuth authentication [5]. To begin the process we need to register our client application with Twitter, create a new application and once you are done we should have our consumer token and secret which are stored in JSON file to makes requests to Twitter for specific data.

C. Tweepy Module

Tweepy is an easy-to-use Python library for accessing the Twitter API [2]. It is a Twitter streaming API used to download twitter messages in real time. It is useful for obtaining a high volume of tweets, or for creating a live feed using a site stream or user stream. This Streaming API only sends out real-time tweets. Using the streaming api has three steps -

1) Create a class inheriting from StreamListener
2) Using that class create a Stream object
3) Connect to the Twitter API using the Stream

These are the key components of tweepy which monitors tweets in real time and catches them. Tweepy also uses OAuth

for authentication and follows the same steps as mentioned in Twitter module.

D. Sentiment Analysis Module

The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's sentiment towards a particular topic, product, etc., is positive, negative, or neutral. We are using a python library named TextBlob for processing textual data [3]. It provides a simple API for diving into common natural language processing (NLP) tasks such as sentiment analysis and more. After evaluating the sentiments and their properties, it returns a namedtuple of the form Sentiment (polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. This helps us to categorize the data into positive, negative and neutral sentiments where positive is represented by 1, negative is represented by -1 and neutral by 0.

E. Plotting Module

Bokeh is an interactive python visualization library that targets modern web browsers for presentation [8]. Its goal is to provide elegant, concise construction of versatile graphics, and to extend this capability with high-performance interactivity over very large or streaming datasets. Bokeh can help anyone who would like to quickly and easily create interactive plots, dashboards, and data applications. The architecture of Bokeh is such that high-level "model objects" (representing things like plots, ranges, axes, glyphs, etc.) are created in Python, and then converted to a JSON format that is consumed by the client library, BokehJS. We are using various Bokeh modules such as bokeh.models, bokeh.events and bokeh.io which aid in creation of widgets, their associated events and displaying them on the browser [Refer Fig. 1]. Our project consist of various widgets such as textbox, dropdown menu, submit buttons , google maps etc. These fields are displayed on the browser when we run Bokeh server on localhost. Since our project requires plotting data on world map, we are using google map widget of Bokeh to display co-ordinates of tweet. The Google map uses Google map API key for authentication. To make our own key, we need to create an account and generate your own unique google api key. This key is used to track the daily uses of google map API. We extract the data from the csv file and this data is plotted on the map using the above module [Refer Fig. 2].

III. DATA STORAGE

After evaluating the tweet data obtained from Twitter API and Tweepy API, we compute the sentiments for those tweets which consist of location co-ordinates and store the data obtained in comma separated values files i.e. a csv file. This csv file consist of 4 columns namely text, latitude, longitude and sentiment. This data is coined from the above 3 modules mentioned above.
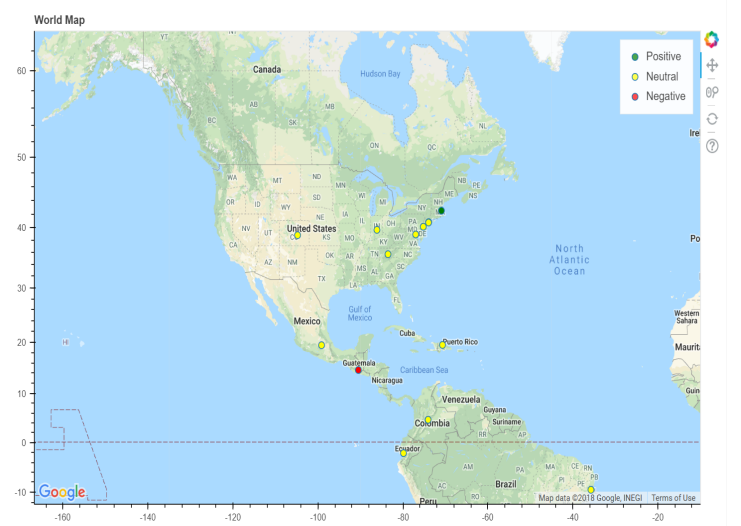

Fig. 1: Bokeh Widgets

Fig. 2: Live Sentiment Plotting

IV. ALGORITHM

For our project, we accept two user input, one in the form of a word which can be a trending topic or a hash tag and the other as the timeframe of the tweets. To ensure we analyze each tweet exactly once and extract result based on unique data, we fetch the first tweet id based on the given timeframe entered by the user and the last tweet id available till date. These id's are called since_id and max_id. Hence, the tweets are returned with a tweet ID greater than since_id and less than max_id. Using our user defined function get_tweet_id we extract these id's. However, there is a limitation on the number of tweets retrieved per request. Hence, we need to fetch tweets in the batch of 100 tweets per request with the help of our user defined function fetch_tweet which internally calls Twitter search API's mentioned above in Twitter module. Once we get the tweet data, we need to filter it and discard those tweets which do not consist of location co-ordinates. On this filtered data set, we then perform sentiment analysis using our user defined function check_sentiment. This internally makes use of the above mentioned sentiment analysis module. The next

step is to store the data in the CSV file.

While this data is being fed to the plotting module, we fetch tweets from another data source. These are the real time tweets which are being fetched using the Tweepy module. This tweepy stream listens to the live tweets at an interval of 5 seconds for 10 times. Once we fetch the tweet data, we then filter them based on the location co-ordinates and perform the sentiment analysis. The result is then stored in the same CSV file mentioned above.

V. LIMITATIONS AND THEIR RESOLUTIONS

During the course of our project execution, we encountered a couple of hurdles. The biggest hurdle for our project was fetching large amount of data without interruption. Second limitation was the availability of location co-ordinates for tweets, as it is not necessary that every user has enabled their location settings. Apart from these two limitations even the API's had some restrictions of their own. Twitter Standard Search API has maximum number of tweets returned per page set to 100 tweets per request. This API also has a strict rate limit on how many request you can send given a time limit and how many tweets you can get access to for each request. The default rate limit is 15 minutes. Hence, we will encounter an exception after 15 minutes. Therefore to overcome these limitations, we needed to execute our program for as long as possible to fetch the maximum number of tweets without interruption. After a lot of trial and error, we deduced that we can retrieve maximum 17500 tweets in one go without any interruptions.

VI. CONTRIBUTIONS

We strongly believe that the strength of a team is equivalent to the strength of the weakest member of the team. Hence, it was important that there should be proper distribution of work amongst us to ensure that everyone gets equal opportunity to learn, grow and enhance their skill set. Hence, as a team we decided that all the work would be distributed fairly amongst us. We equally distributed the coding part by doing pair programming, divided the slides for presentation and also distributed the Latex write - up which includes formal documentation and report.

VII. CONCLUSION

To conclude, our project provides a geographical visualization of the trending topic along with different sentiments i.e. positive, negative, neutral on the World Map (color coded) to enable users to observe people's sentiments about a particular topic in different parts of the world. Social networking websites and related tools allow companies to speak with, ask questions, answer questions and overall interact with their customers as never before. This projects will act as a value added service and will be beneficial for those who want to know the sentiments for a given topic. For example, this can help small business or organizations grow by getting an overview about the sentiments regarding their new or existing products, how well is it doing or are there any complaints regarding the same etc. It can also help politicians with their campaigns status and identify which area's would require improvement in regards to their campaign.

REFERENCES

[1] https://developer.twitter.com/en/docs/tweets/search/overview
[2] http://tweepy.readthedocs.io/
[3] http://textblob.readthedocs.io/en/dev/
[4] https://www.telegraph.co.uk/technology/0/social-media-did-grow-quickly/
[5] http://socialmedia-class.org/twittertutorial.html
[6] https://whatis.techtarget.com/definition/social-media
[7] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[8] https://bokeh.pydata.org/en/latest/