Sentiment and Topic Analysis on Social Media

Harshit Garg
hgarg1@binghamton.edu
SUNY Binghamton

Kasturi More
kmore4@binghamton.edu
SUNY Binghamton

Sagar Vishwakarma
svishwa2@binghamton.edu
SUNY Binghamton

ABSTRACT

Social media platforms like Facebook, Twitter, Reddit, YouTube, and other allow users to share their ideas and online content and interact with each other through the virtual networks and communities.

Data analysis plays an important role in handling and managing these social media platforms starting from creating a profile for a user, keeping track of shared data, likes and dislikes for a particular post to advertisements shown. On Facebook, we can collect data of increases in followers, numbers of likes and dislikes or number of shares. From twitter, we can collect data of the retweeted post and on Instagram, hashtag usage and engagement rates. There can be much information collected from this like demographic information, sentiments for the trend and many more. With technology's increasing capabilities, sentiment analysis is becoming a more utilized tool for businesses. Social media monitoring tools use it to give their users insights about how the public feels in regard to their business, products, or topics of interest and can help you see how positively or negatively your brand is perceived on social media, based on the tone of mentions.

The goal of this project is to find some data driven insights of the trending topics on platforms like Twitter and Reddit. Where with the help of just the popular or trending or any keyword, sentiment analysis will be there in the conclusion.

1 INTRODUCTION

Social media, such as Twitter and Reddit, provide exciting opportunities that can "open up a new era" of social science research. These new communication platforms afford the ability to examine social data on a variety of topics, on a massive scale, and over short periods of time. Despite recent and growing interest in using Twitter and Reddit to examine human behavior and attitudes, there is still significant room for growth regarding the ability to leverage these data for social science research. Twitter provides a convenient source of data on users' opinions, interactions, and reported behaviors. One popular application of Twitter and Reddit data is sentiment analysis. We have collected all the posts and its related metadata for selected trending topics in the World. Topics we have selected for this project are: US Elections, Covid, Work from home, h1b and stocks. After performing promising analysis on this collected data, we will be able to perform sentiment analysis on corresponding topic. For example, for US elections, we will perform political analysis and election forecasting to predict political outcomes, track changing tides in political opinion and examine intention to not vote. For Covid, we will be analysis how people are reacting and handling the outbreak of corona virus. For collecting twitter data, we have used Twitter stream API and for collecting Reddit data, we have used Reddit API. No high-level library is used such as Tweepy, PRAW and scrapy.

2 IMPLEMENTATION

We are collecting text data from tweets and reddit posts to analyze how the public feels and responds on a topic and find out how we can use this data to understand general notion about the topic.

1. Accessing credentials: First, we are accessing credentials from 'credentials' file using ConfigParser. Accessed credentials will be used in a request header. Credentials file has details of "api key", "api secret key", "bearer token", "stream bearer token", "access token" and "access token secret" for Twitter. Along with above mentioned details for Twitter credentials, this file also consist of "username", "password", "app name", "public key" and "secret key" for Reddit.

2. Accessing trending topics: We have created a file named "top topics" to store list of all trending topics. Before accessing data from Twitter amp; Reddit, we first retrieve these topics from file and store those topics in an array.

3. Accessing Twitter Data: To collect twitter data we have implemented python script. We are accessing the (publicly available) data from twitter using the Twitter Streaming API as well as REST API. We have accessed these twitter data using the persistent HTTP GET connection and used OAuth 2.0 Bearer token authentication. From Twitter Stream API, we are taking randomly selected 1tweet from the real time. Tweets delivered through sample streamed do not count towards the monthly count cap.

We are connecting to a Twitter end point by sending GET request with search URL and authorization header and we are storing each response line in a JSON format.

To convert these response lines into JSON-encoded data, we have used the loads method from the json module. Whenever we request a tweet data, Twitter API will by default send back ID and Text for tweets. However, to get some more fields of tweets, we need to add object type and field name in a search URL. For example, https://api.twitter.com/2/tweets/sample/stream?tweet.fields=created_at,lang After receiving all the raw data in JSON format, we iterated through that data and stored it in a dataframe which is used to format and add additional data which will be later converted into a json format for storing in the database.

4. Accessing Reddit Data: We have implemented python script for reddit data. We are accessing the data from reddit is done using the Reddit API. We are accessing these reddit data using the GET request to URL https://www.reddit.com/search?q=election2020 To retrieve the data in a JSON-encoded format, we added .json to the end of this request (i.e., https://www.reddit.com/search.json?q=election2020). However, there is an issue with this request - by default, it was only giving us 25 posts. As mentioned in an API documentation, this request is a listing, therefore it takes the parameters after, before, limit, count, and show. There is also an additional parameter t, which limits the time of the posts to show. So, to get the posts from

CS580N'20, Sep 2020, Binghamton, NY, USA

one day, we changed our request to https://www.reddit.com/search.json?q=election2020&amp;t=day. Additionally, if we want a specific number of posts (let us start with one), we add the limit parameter to get https://www.reddit.com/search.json?q=election2020&amp;t=day&amp;limit=1. The two most important rules set up by Reddit for accessing the site are: a) We need to create a unique UserAgent string - reddit requires that the UserAgent string contains username, b) we need to limit the number of requests we send to less than 30 a minute (i.e., one every two seconds). As we noticed from the documentation, reddit only allows to pull 100 posts at a time from the search board of a subreddit. We are using unique identifier which is stored in the name key for each post to track the last post we get from each request and then starting the next request after this post. We simply coupled this with the after parameter in a loop to get the number of posts we want. In other words, we sent a request for the first 100 posts, obtained the identifier of the final post in that request, and asked that the next request of 100 start at the first post after this post. In our case, we have extracted 1000 posts. As part of extracting the data, we kept only the content assigned to the children key in the first dictionary. This makes it possible to put all the separate requests together in a collection (in this case, a list). To make sure we do not accidentally exceed reddit's request limit of 30 requests per minute, we have used the sleep method from the time module to place a pause of 2 seconds in between each iteration of the loop. Now that we have our data in raw JSON format, we simply used another loop to extract the desired information from each post. In our case, we have decided to get the date of posting, title, number of comments and URL. Finally, we have created pandas DataFrame using these lists of results

3 STATISTICS

From Reddit, we are getting 1000 records and those records are consuming 0.45 MB space in the memory and the execution time for this is 671 sec. From Twitter stream, we are getting approximately 125k records and it is consuming 64 MB space for the collected data for one day data and the execution time for this is around 4465 sec. From Twitter, we are also collecting data based on the topic and getting roughly 500 records and those records are consuming 0.36MB in the storage for one day and the execution time for this is around 3.181sec. Similarly, we are expecting the same range of the data every day for the next 7 days for this project's implementation

4 CHALLENGES

One of the main challenges in this project is numerous differences in Twitter data and Reddit data. Along with collecting data from two different platforms, its also challenging to normalize such data to common grounds. In addition, it is difficult to gather demographic data from text-based posts.

REFERENCES

1. Yang Zhang. Language in Our Time: An Empirical Analysis of Hashtags. https://yangzhangalmo.github.io/papers/WWW19.pdf

2. A comparative analysis. https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.2016.14505301151

3. A Comprehensive Analysis of Twitter Trending Topics. https://arxiv.org/ftp/arxiv/papers/1907/1907.09007.pdf

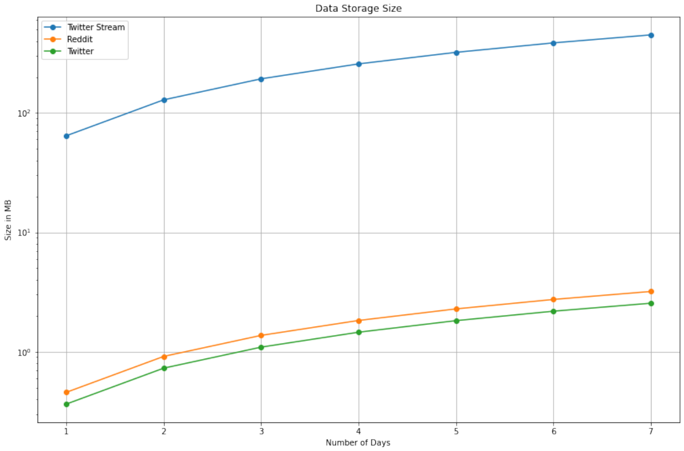

Figure 1: Data Storage Size