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.

We are collecting the data useful and required for our research. As we are performing sentiment analysis on 5 trending topics, we are collecting all the posts from Twitter Reddit for those topics. Topics we have selected for this analysis are 1) US Elections, 2) Covid, 3) Work from Home, 4) H1B, 5) stocks. The goal of this project is to find some data driven insights of topic on platforms like Twitter and Reddit.

Collected data will be analyzed with various computational methods to determine public's reaction or thoughts about a topic. For example, for US elections, we will perform political analysis and election forecasting about how people feel about elections and analyze positive and negative sentiments towards a topic.

Twitter API Reddit API is used to collect real-time data. Data collection has been implemented in Python.

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. Sentimental analysis is the process of computationally determining the opinion or attitude of the writers as positive, negative or neutral. In many fields like business, politics and public actions, determining the sentimental analysis is very important. Considering business, it is very useful to understand the customer's feelings in order to develop their company. Next in politics: It can be even be used to predict the election results. Our project focuses on predicting the general sentiment polarity of the reactions on a topic based on posts on Reddit Twitter.

People post their opinions, view on a topic on social networking sites. Thus, we have collected and still collecting all the posts and its related metadata for selected trending topics in the World. Topics we have selected for this project are mentioned above. After performing promising analysis on this collected data, we will be able to perform sentiment analysis on corresponding topic. 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.

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

Second phase of the project focuses on designing and executing measurements and analysis experiments on collected data. This paper proposes how collected data will be used, what analysis will be performed and methodologies we will use for the same.

2 RESEARCH OBJECTIVE

As mentioned earlier, we will use Twitter and Reddit data. Twitter Reddit data are abundant and relatively easy to access. All published content for unprotected user is available for view to all web users that means it is public data. Users do not need to issue approval for researchers to use their profile information. Thus, collecting data from Twitter and Reddit and performing analysis on these data is legal. We are collecting Id, text, and timestamp from the twitter while from reddit we are collecting id, current date, postdate and the text for the analysis. We have pre-defined the topic "US Election", "COVID", "Work from Home", "H1B" and "Stocks" for the sentiment analysis. Based on the English language we will be filtering the tweets and subreddit post for further use.

We will try to answer following research questions:

(1) What people think about each mentioned topic and how many people are positive, negative, and neutral about it on both the social media platform separately.
(2) Comparison of the analysis of each topic on both the platform.
(3) We will try to run the sentiment analysis on the Daily, Weekly and Monthly basis and try to understand how the sentiment is changing with time for the mentioned topics.

3 METHODOLOGY

Methodology we will be following to answer our research questions mainly comprises of 4 steps namely Data Collection, Data Cleaning, Data Analysis and Analysis Results.

3.1 Data Collection

Data collection is nothing but tweets extraction and Reddit posts extraction. Twitter maintains multiple free, public options for accessing data. Each method has unique advantages and disadvantages. Our approach to collecting Twitter data involves access through Twitter's streaming API, which provides us a query able sample (1 percentage of all content) of tweets created in real time. Advantages of using this version of the API include its speed and the volume of data available.

Reddit offers a simple Application Programming Interface (API) to facilitate data collection from Reddit. We implement the Reddit API and a scraper code in Python 3.6. For every posting, we collect the posting date and text. Data storage is implemented with MongoDB. All the implementation details used for twitter and reddit has been provided in Project 1.

3.2 Data Cleaning and Preprocessing

A tweet or a post contains a lot of opinions about a topic which are expressed in different ways by different users in different languages. Preprocessing of tweet include following points

- One of the main challenges in data collection from Twitter and Reddit is unpredictable content such as bots and tweets/posts unrelated to a specific query may find their way into the data and skew results. Thus, the removal of irrelevant tweets is important. we clean irrelevant tweets/posts using a list of keywords created using a combination of data-driven techniques, including topic names.
- Removing all the tweets and posts in languages different than English.
- Removing all URLs, @Usernames and stop words like he, she, it, they, or and many more
- Normalizing the data. As we are collecting data from 2 different platforms, we are going the normalize these data.
- Standardizing date format to mm/dd/yyyy
- Emoticons Replacement. Emoticons are very important in determining the sentiment. So the most common emoticons will be replaced by their polarity by seeing the emoticon dictionary and others will be removed

3.3 Data Analysis

After cleaning and preprocessing raw data, we will be using machine learning technique to classify a tweet or a post based on its polarity. We are planning of using Naive Bayes method for the classification. This technique is very easy and fast to predict the class of data set.

3.4 Analysis Results

Once we run all the data analysis, we will conclude polarity and public's reaction towards a particular topic. We will be using Matplotlib library for the projection and comparison of the graphs of analyzed data. We will also compare results obtained from Twitter and Reddit data. We can use this sentiment analysis performed to identify for how much time particular topic is in trending and people give positive, negative or neutral reactions on it.

4 DATA REQUIREMENTS

From Reddit, we are getting 1000 records daily and those records are consuming 0.45 MB space in the memory. From Twitter stream, we are getting approximately 125k records daily and it is consuming 64 MB space. We will continue collecting these data every day for the next 10 days for this project's implementation.

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

4. Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. https://www.researchgate.net/publication/282878896_Using_Twitter_for_Demographic_and_Social_Science_Research_Tools_for_Data_Collection_and_Processing

5. Sentiment Analysis of Twitter Data: A Survey of Techniques. https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf

6. Predicting sentiment of comments to news on Reddit. https://esc.fnwi.uva.nl/thesis/centraal/files/f668331765.pdf