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. Social media has become a treasure of information.

Sentiment analysis plays an important role in decision making. It is also very useful in recommender system. With the help of polarity defined, one can study various trends and popularity of certain things which will be eventually helpful in making important decisions.

As we performed sentiment analysis on 5 trending topics, we have collected the posts from Twitter and 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 was to find some data driven insights of topic on platforms like Twitter and Reddit.

Twitter API and Reddit API is used for collecting real-time data. Collected data was analyzed using VADER(Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool to determine public's reaction or thoughts about a topic and for calculating semantic orientation of each posts. For example, for US elections, we have performed statistical analysis to see how people feel about elections and analyzed positive and negative sentiments towards a topic. We have also performed comparative analysis to see how polarity varies with the different social media platforms.

1 INTRODUCTION

Social media, such as Twitter and Reddit, "opens up a new era" of social science research by providing exciting opportunities. These new communication platforms afford the ability to examine social data on a variety of topics, on a massive scale. 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 identifying and categorizing the opinion or attitude of the writers as positive, negative or neutral by analyzing the text. 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 reviews and feelings in order to develop their company or product. [7] 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 and Twitter.

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

People post their opinions, view on a topic on social networking sites. To categorize posts into 3 sentiments that is positive, neutral and negative and perform analysis based on that, various steps needs to follow that are data collection, data pre-processing, sentiment classification and analysis. We have collected raw data from Twitter and Reddit. Data collection for selected five trending topics was performed from 22nd November to 28th November. 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.

As we have collected data from two different platforms, it is important to standardize the data to perform analysis. Thus, data pre-processing has been performed. On these pre-processed data, we have used VADER sentiment analysis tool to categorize a post into positive, neutral or negative sentiment. VADER uses lexicon based sentiment classification approach. [3] After sentiment classification, we have performed topic-wise comparative analysis to understand trends and how public opinion changes over time. How the trend or opinion change is different on both the social media platforms.

Our project focuses on answering 3 research questions:

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

The report structure is as follows. In section 2, we present background and related work to describe the problem domain we are working on followed by dataset used in section 3. Section 4 is description of the methodology applied and section 5 presents results obtained after analysis. And finally last section describes conclusion and future work.

2 RELATED WORK

Previous research and projects on sentiment analysis uses various approaches for opinion mining. Okanohara and Tsujii [4] has performed binary sentiment classification. They have conducted experiments on book reviews on Amazon.com having five point scale rating with text. Leimin and Catherine [5] believe that most of the previous work on sentiment analysis in NLP focuses on text i.e. tweets or news articles. However, human language is multimodal. Thus, they have conducted multimodal sentiment analysis.

Fabio and his team [1] has proposed a sentiment analysis classifier, named Senti4SD, which exploits a suite of lexicon-based, keyword-based, and semantic features for dealing with the domain-dependent use of a lexicon. Some of the early results on sentiment

analysis of Twitter data are by Go et al. Go et al [2] use distant learning to acquire sentiment data. They use tweets ending in positive emoticons like ":)" ":-)" as positive and negative emoticons like ":(" ":-(" as negative. They build models using Naive Bayes, Maximum Entropy and Support Vector Machines (SVM), and they report SVM outperforms other classifiers.

3 DATASET

We have collected raw text data from tweets and reddit posts using platform APIs 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. Data collection was done during 21st November to 27th November. First 2 steps in data collection are common for both the social media platforms. First step is accessing credentials from a file storing usernames, passwords, public keys, API keys required to access data through API. Second step is accessing all the trending topics we have selected.

3.1 Twitter Data

For collecting Twitter data we have used Twitter stream API and have implemented Scraper code in Python 3. We are accessing the (publicly available) data from twitter using the Twitter Streaming 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 1 percent tweet from the real time. Tweets delivered through sample streamed do not count towards the monthly count cap. 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 have added object type and field name in a search URL. We have stored these raw data in JSON format in mongoDB database. Over a span of week we have collected almost 80 lakhs records through Twitter stream API and 10000 records for Twitter API.

3.2 Reddit Data

For collecting Reddit data we have used Reddit API and have implemented Scraper code in Python 3.

We are accessing these reddit data using the GET request to URL https://www.reddit.com/search?q=election2020 and retrieving it in JSON format. we have stored date of posting, text, and URL for each reddit post. We have stored these raw data in JSON format in mongoDB database. Over a span of week we have collected almost 21000 records through Reddit API.

4 METHODOLOGY

Methodology we have followed to answer our research questions mainly comprises of 4 steps namely Data Collection, Data Cleaning, Sentiment Classification and Analysis.

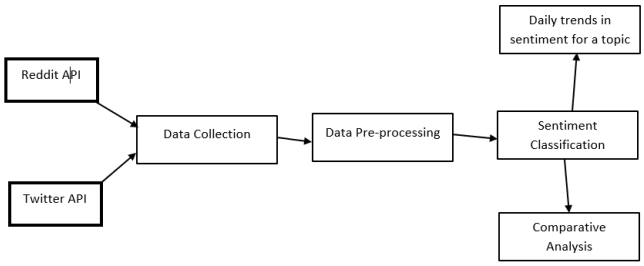

Figure 1: Project Methodology

4.1 Data Collection

Data collection is nothing but tweets extraction and Reddit posts extraction. Twitter maintains multiple free, public options for accessing data where each method has unique advantages and disadvantages. Our approach to collecting Twitter data involves accessing tweets 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 have implemented 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 previous section.

4.2 Data Pre-processing

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.
- Standardizing date format to mm/dd/yyyy

4.3 Sentiment Classification

After preprocessing raw data, we have assigned a polarity to each record. For sentiment analysis we have used VADER sentiment analysis tool to categorize a post into a "Positive", "Negative" or "Neutral" sentiment.

VADER is a rule-based sentiment analysis engine which calculates pos, neg and neu scores for each record. These pos, neu, and neg scores are ratios for proportions of text that fall in each category. [6]

The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules. This score is then normalized between -1 to +1. -1 is considered as most extreme negative while +1 is most extreme positive. Compound

score is very useful for unidimensional measure of sentiment for given sentence. If compound score is >= 0.05 for a sentence then it has a Positive sentiment. If compound score is <= -0.05 for a sentence then it has a negative sentiment. Otherwise a sentence has a Neutral sentiment.

4.4 Analysis

After assigning a sentiment polarity score to each record, we have performed various comparative analysis on the data. Details for these analysis results are described in next section.

5 OBSERVATIONS AND RESULTS

- Twitter vs Reddit COVID result: From the graph, we can say that from Nov 21st to Nov 25th comparison of sentiments for twitter and reddit are following same trends. While, from Nov 25th to Nov 27th percentage of positive sentiments is increasing for the reddit data and the same is decreasing on the twitter data. Negative sentiments are decreasing for the reddit data and increasing for the twitter data. Overall Neutral sentiment for this period is slightly increasing on both the platform.

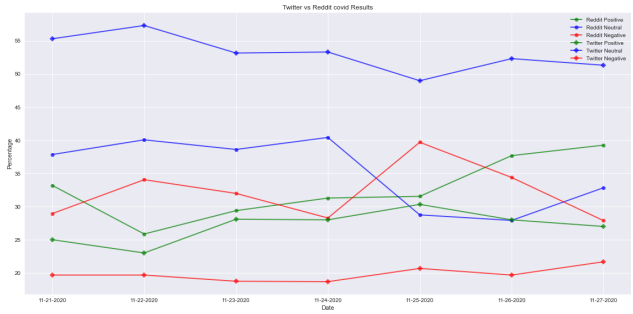

Figure 2: Twitter vs Reddit covid Results

- Twitter vs Reddit h1b result: Form the graph, we can say that comparison of sentiments for twitter and reddit are not following same trends, rather it is changing everyday for both the platforms. For example, positive sentiments are increasing from 0

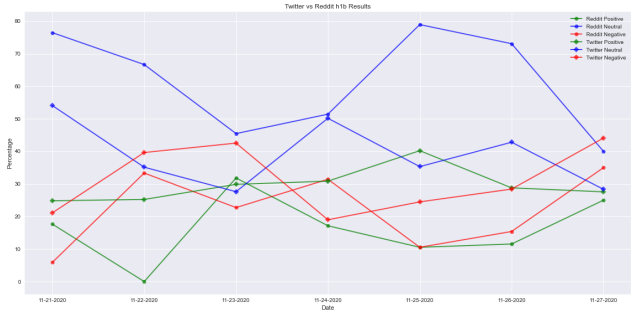

Figure 3: Twitter vs Reddit h1b Results

- Twitter vs Reddit stock result: Form the graph, we can say that comparison of sentiments for twitter and reddit are following opposite trends for most of the time. For example, for the first day Nov 21st to Nov 22nd positive sentiments are decreasing for the twitter and it is increasing for the reddit data. Similarly, most of the time if the positive sentiments are increasing for the twitter data then the same is decreasing for the reddit data and the same trends we are observing for the negative sentiments also.

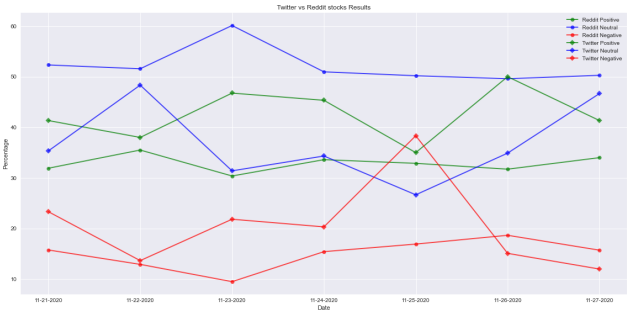

Figure 4: Twitter vs Reddit stocks Results

- Twitter vs Reddit US Election result: Positive sentiments have a drastic change in twitter data almost everyday for the twitter data while there is not much change for the same for the reddit data. Similarly, negative sentiments of twitter also have many changes compare to reddit data. Reddit's negative sentiment is either flat or decreasing slightly. Neutral sentiments are almost changing slightly for the first 3 days while next two days it has more changes for both the platforms.

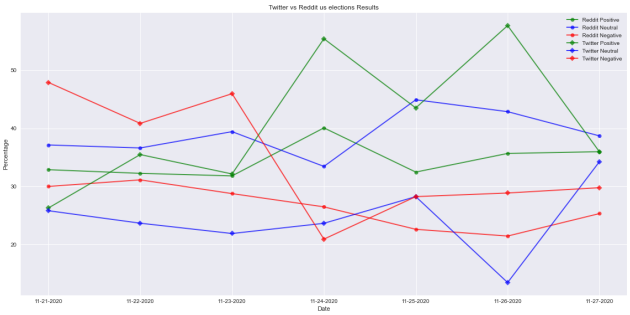

Figure 5: Twitter vs Reddit us elections Results

- Twitter vs Reddit work form home result: For the first 4 days, change in positive sentiments on twitter and reddit are following the opposite trend. If twitter trend increases, then reddit trend decreases. While for the next two days twitter has a flat graph and reddit has slight increase in positive sentiments. Similarly, neutral sentiments are also following the opposite trends for almost all the days for both the platforms. Negative sentiments of reddit data has less change everyday compare to the twitter's negative sentiments for the greatest number of days.
- Twitter stream results: Over a period of 7 days data collection, graph for all the positive, negative and the neutral sentiments are almost flat and does not has a big change in any of the day. We are getting roughly 65
- Twitter stream data collection results: From the graph, we can say that every day we are getting almost same kind of

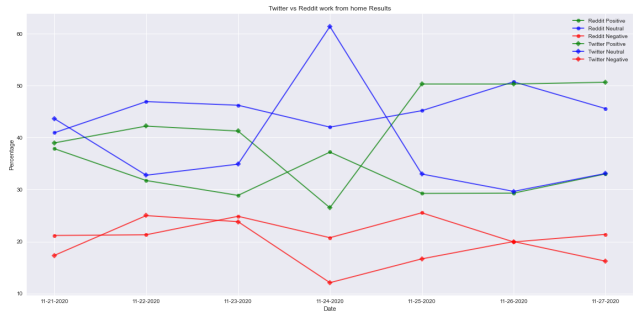

Figure 6: Twitter vs Reddit work from home Results

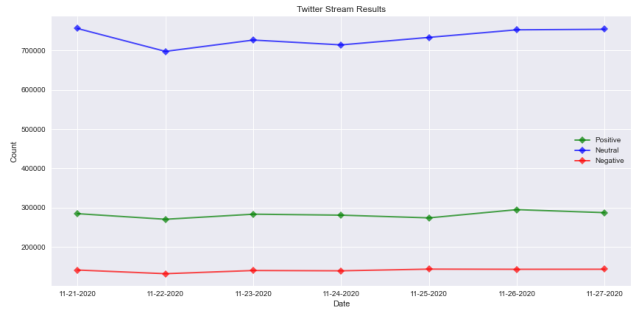

Figure 7: Twitter Stream Results

count graph pattern. we have a very interesting result of tweeter steam count with the time and we can see maximum number of tweets around 3 am and 3 pm. Total tweets collected for any day for one hour is between 100k to 160k.

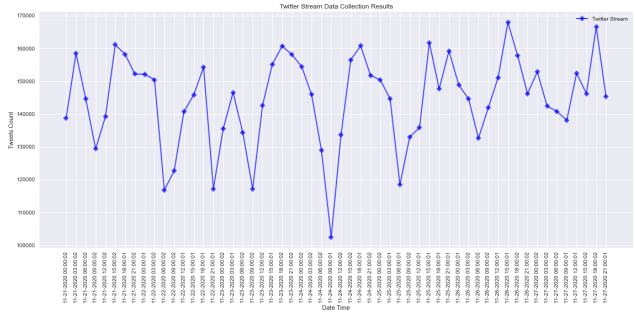

Figure 8: Twitter Stream Data Collection

6 CONCLUSION AND FUTURE WORK

We try to build sentimental analysis system with the help of Vader's sentiment analysis for the below mentioned five topics. 1) Covid 2) H1 b 3) Stocks 4) US election 5) Work from home

We are getting more number neutral sentiments than the positive and the negative for almost all the graph as there are other languages text from tweeter and reddit for the period of Nov 21st to Nov 27th. In our project we can also add the any topic at any point of time and collect the data and do the sentimental analysis. In the future work, this analysis can be done for the longer period. Also, we have observed more data from reddit compare to the tweeter for a particular topic. why we are getting more data from reddit? We are also getting many different language texts from both the platform and how this other language text can be converted to the English and can be used for sentiment analysis. These are some of the areas which can be worked on the future work.

	Twitter	Reddit	Twitter Stream
Positive	36.414	32.894	24.427
Neutral	39.751	43.037	63.399
Negative	23.833	24.067	12.173

Above table depicts the percentage polarity trend over various platforms.

REFERENCES

[1] Fabio Calefato, Filippo Lanubile, Federico Maiorano, and Nicole Novielli. 2018. Sentiment polarity detection for software development. *Empirical Software Engineering* 23, 3 (2018), 1352–1382.

[2] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision.(2009). (2009).

[3] Vishal Kharde, Prof Sonawane, et al. 2016. Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971* (2016).

[4] Daisuke Okanohara and Jun'ichi Tsujii. 2005. Assigning polarity scores to reviews using machine learning techniques. In *International Conference on Natural Language Processing*. Springer, 314–325.

[5] Leimin Tian, Catherine Lai, and Johanna D Moore. 2018. Polarity and intensity: the two aspects of sentiment analysis. *arXiv preprint arXiv:1807.01466* (2018).

[6] W. Weerkamp and M. D. Rijke. 2012. Predicting sentiment of comments to news on Reddit.

[7] Yang Zhang. 2019. Language in Our Time: An Empirical Analysis of Hashtags. (2019). arXiv:cs.SI/1905.04590