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. [3] 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. [2] 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 DASHBOARD

2.1 Proposed Analysis

We will be displaying individual and comparative analysis of Twitter data and reddit data for selected 5 topics. In an individual analysis, selected a date range with a platform for example Twitter, all the topics comparison will be shown based on selected sentiment. This is to show how each topic's sentiments are varying each day compared to others.

In comparative analysis, comparison of a particular topic on 2 different platforms i.e. Twitter and Reddit will be displayed for all sentiments. Comparative analysis is relative comparison on both platforms rather than one to one comparison.

For Twitter stream, given a date range sentiment for all data collected will be compared.

We will also plot data collection results for given date range.

2.2 Tools

Tools we will be using for dashboard creation – Dash, numpy, pandas, matplotlib.

Dash is a productive Python framework for building web analytic applications. [1]

Written on top of Flask, Plotly.js, and React.js, Dash is ideal for building data visualization apps with highly custom user interfaces in pure Python. It's particularly suited for anyone who works with data in Python.

Through a couple of simple patterns, Dash abstracts away all of the technologies and protocols that are required to build an interactive web-based application. Dash is simple enough that you can bind a user interface around your Python code in an afternoon.

Dash apps are rendered in the web browser. You can deploy your apps to servers and then share them through URLs. Since Dash apps are viewed in the web browser, Dash is inherently cross-platform and mobile ready.

Dash is an open source library, released under the permissive MIT license. Plotly develops Dash and offers a platform for managing Dash apps in an enterprise environment.

REFERENCES

[1] 2017. Multilayer Perceptron. https://dash.plotly.com/layout. (2017).

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

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