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 IMPLEMENTATION

We will be collecting tweets and posts and 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.

To analyze the tweets and posts about a topic, we are planning to use the WEB APIs provided by Twitter. This can be achieved by creating a python application that periodically runs the APIs fetches the data and stores in the database. For data storage we have decided to use the MongoDB. For integration of the python application to the MongoDB a wrapper interface like PyMongo can be used. For the periodic job scheduling, a scheduler like CronJobs can be used.

The Twitter API lets you access and interact with public Twitter data. We will be using the Twitter Streaming API to connect to Twitter data streams and gather tweets containing keywords, brand mentions, and hashtags, or even collect tweets from specific users. The Standard Search API lets us get historical tweets published up to 7 days ago. Alternatively there are historical search APIs (like Historical PowerTrack and Full-Archive Search), that can collect tweets from as early as 2006. Similarly there are APIs like PRAW for reddit. Though the mentioned search APIs allows search access to tweets and posts up to a week old, the script to collect data must run more frequent as tweets and posts can be removed quickly through moderation. The frequency of running a script to call this API will be decided after running a small test and will be based on the amount of text data and other additional information about the posts.

The topic will be selected before running the script and can be any topic we want, we will be storing the data daily and the stored data will be divided into weekly, monthly and historical. The process will move the data from the weekly table to monthly at the end of the week, similarly the data collected in monthly will be moved to historical at the end of the month. This will help us restrict our analysis to a specific time window.

Sentiment Analysis will be used to find out the volume of positive, negative or neutral responses on social media platforms, and along with statistical analysis on the additional data about the post.

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. How to Scrape Reddit with Google Scripts. https://www.labnol.org/internet/web-scraping-reddit/28369/

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