Capstone Project

Rohit Savalgi May 2023

# Introduction

Many people start their day with Twitter. It’s where some people collect ideas, opinions, and even so-called” facts” from the news, but one thing that is common between all users is the negative emotions that circulate the app. With one negative tweet many more appear, creating a storm of” haters” on the internet. But what if the app was designed to intentionally help people that are tweeting things that they view negatively, almost like they had a place to go to understand different viewpoints and understand the positives and the negatives? With this project I intend to create a model that will help users look at all viewpoints of a certain topic by searching through all tweets dedicated to the topic, and then separating it into the the positive or negative sentiment. This will then display the related tweets and give the user the option to browse through the many other opinions on Twitter without giving a bias view on the topic. Tweet analysis has been done before, but not with the same intention as this model. This model is focused on giving people the whole story from credible sources all over the ”Twittersphere” while other models simply give related searches based on keywords without any real intention or motivation.

# Literature Review

There are many twitters sentiment analysis done throughout the internet, but not all are specific or meaningful. On the dataset being worked on there is code that analyzes tweets with deep word encoding, searching, and recognizing. But with this no actual intentional analysis is being done. The work is done by Pavan Sanagapati with his work on knowledge graphs and NLP (Natural Language Processing). The code is amazingly done and can be used well for parsing through words and names and comparing them to other entities.

Another similar sentiment analysis has been done before by analyzing hatred tweeted out and what its centered around. The model layout will be similar in this project, but mine will focus on connection and help through sentiment analysis tweets rather than hatred. With Python and a sentiment analysis dataset the author, Mustanger, easier parsed through and uncovered the hatred tweets topics.

# Dataset

Twitter Sentiment Analysis Dataset: Train and Test Files

GitHub: HTTPs://github.com/GitRhoeatyHub/492Capstone

# Exploratory Data Analysis

For now, the datasets being used consist of only three variables. An ID for the tweet, what the tweet says, and the sentiment of the tweet (0-1, positive- negative). I intend on using all three variables, but as the project goes on new pieces of code will be needed, so more tables and data will be required for testing and training.

# Methods

The data is thin but consists of many values; a knowledge graph is one specific way it will be organized. Based on that the topics of the tweets will be found with the closeness of the knowledge graph. This process will be remarkably similar to clustering, with this we are getting the most related hits. As well as clustering, randomness will be used to shuffle the knowledge graph occasionally to see any things that may have been missed.

A pro of this strategy is that clustering the similarity of tweets which I am looking for will be much easier, but with the extensive list of tweets it may also be a con. A pro also is that clustering is a hands-off process that is mostly handled by machine and will calculate less error. Another con is the possibility of poor descriptors for clustering, the English language is big and the tweets that may be attracted to each other based on the word won’t always be the most influential or one that can help out hypotheses. All this will be considered.

# Rationale

The methodology used is painstaking as there is a lot of pre-pre-processing. In order for the clustering to work many of the tweets had to become filtered by taking out stop words, useless words, fixing spelling, or creating a better general idea for the tweet. With this preprocessing done, another sentiment analysis is done, but this way we try to figure out how sentimental and in which direction the tweet is going, positive or negative. This along with our frequency identifiers we are able to tell the meaningfulness of the tweets in how they are valued within the tweet and comparatively towards other tweets. After this processing is done, we simply search for a word that we want to relate to. Then with the provided a code a list of tweets will pop up showing the most related tweets in positive or negative sentiment based on what the user wants.

