Data Collection System Measurements and Analysis Proposal

Omer Feyzullahoglu
Binghamton University
Binghamton, New York, USA
ofeyzul1@binghamton.edu

ABSTRACT
In this paper, I describe the road map that I will follow while developing measurements and analysis system for social media data science project.

Keywords
Proposal, data collection, Twitter, Reddit, python, Vader, text blob, Jaccard index

1. MOTIVATION
Data Science is without a doubt one of today's most prominent study and application fields. The number of people interested in learning Data Science, which is by its very nature an interdisciplinary field, is growing by the day. [4] We live in a world where data flows in huge amounts and in diversity, and everyone talks about data being a very important source of information. So, what kind of information do we aim to reach from the data? Or what is the motivation that drives people to data science?

People want to get useful information about the past, the present and the future, which motivates them. In the context of this project and from my point of view, the motivation for the experiments I will do in this study comes from the desire to learn new concepts related to data science and to use it together with a different discipline, statistics.

2. RESEARCH OBJECTIVES
The focus of this study is to try to draw meaningful conclusions from the data obtained from social media. In order to achieve this, I determined three different research questions.

According to my observations, people post more negative comments when they are using social media then in the real life. That's why I wondered how much of the data I retrieved from platforms was positive and how much was negative. So, my first research question is “What is the positive/negative expression ratio of the collected tweets and reddit titles?”

For the preliminary exploration of the reddit data, I checked the database manually to see is there any different things apart from I planned. I found out that self-text field generally consists of from links not from texts for first couple of pages of the database entities. So, I wanted to know the ratio of this situation. Therefore, I want to ask my second research question as “What is the text/link ratio of the reddit post’s self-text field?”

On the other hand, generally, there are topics that are more popular than other topics that everyone comments on. I also wonder what these topics are, whether the same topics are discussed in the same week on both reddit and twitter, if there are different agendas, how much this difference is? So, my third research question is “Can we understand agenda of the week and Is there a difference between the agenda of the two platforms?”

3. METHODOLOGY
When I did research to answer the first research question, I realized that I need to do sentiment analysis and there are libraries specially prepared for this analysis. First, I thought of doing it using the "Vader library". Vader Sentiment library is a rule-based sentiment analysis tool that is specially designed for social media.[3] On the other hand, I learned a library called TextBlob after the meeting with instructor. TextBlob is a tool that processes textual data, and it has features such as sentiment analysis, word and phrase frequencies, noun phrase extraction.[2] I have not decided which one I should use.

To be able to find an answer to the second research question, I need to make string operations to identify links inside of the text. I think that I can create a regular expression pattern consists of "http", "www" etc.. Then, I can identify the links inside the self-text area with the help of regular expressions.

In order to answer the third research question, I think I need to create clusters from the data I have. Later, I plan to use the Jaccard index, which is used to measure the similarities and differences between the sets used in statistics.[1] Thus, within the same period, I can determine the differences or similarities in the agenda of two different platforms.

4. DATA SETS
As you know, I created two different collections in the Mongo DB to store the data that I retrieved from Twitter and Reddit. Twitter collection has fields consists of text, public metrics, source, context annotations, possibly sensitive. On the other hand, Reddit collection has fields consist of title, subreddit, self-text, permalink. In the light of this information and my research questions, I think that the data I have is sufficient to conduct the necessary experiments.

5. REFERENCES
[1] Wikipedia. 2021. WikipediA: the Free Encyclopedia. Retrieved from https://www.wikipedia.org/.

[2] Steven Loria, pypi.org, Retrieved from https://pypi.org /project/textblob

[3] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014

[4] Selcuk Yusuf Arslan, 2021. Data science for sustainable development goals. Retrieved November 14, 2021 from https://blogs.eun.org/sdw-blog/2021/05/03/data-science-for-sustainable-development-goals