# Data Collection System Measurements and Analysis Report

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

## ABSTRACT

Devices that can access the internet have now become a part of daily life, utilized by everyone, thanks to significant advances in information and communication technology in recent years. The rise in the number of mobile devices has resulted in a unique trend known as social media, which has consumed human existence in recent years, becoming the focal point of everyday life for many people, particularly the young.

In this paper, I present measurement and analysis system for Reddit's and Twitter's public data. This measurement and analysis system provides both sentiment and similarity analysis.

## Keywords

Social media data, data collection, Twitter API, Reddit API, python, requests, pipelining, sentiment analysis, Jaccard index,

## 1. INTRODUCTION

In recent years social media has evolved and grown to the point that it is now considered mainstream. In 2015-2016, Facebook had 1.59 billion monthly users [7], while Twitter had more than 500 million tweets sent every day. Right now, Facebook has 2.89 billion monthly users and Twitter has 192 million active users. In the light of this information, we see that social media data science is gaining importance year by year.

This project aims to provide introductory information on how to use social media data, what kind of analysis is possible, and what kind of tools can be used when working with social media data. In this direction, three research questions were determined and answers were sought with this project.

According to my findings, people write more rude comments on social media than they do in real life. That's why I was curious in how much of the data obtained from platforms was beneficial and how much was harmful. As a result, first research question is, "What is the positive/negative expression ratio of the gathered tweets and reddit titles?"

During the preliminary exploration of the reddit data, I personally verified the database to see if there were any differences from what I had intended. For the first couple of pages of the database entities, discovered that the self-text field is often made up of links rather than text. So, I was curious about the ratio of this circumstance. As a result, my second research question will be "What is the text/link ratio of the reddit post's self-text field?"

On the other hand, there are some themes that are more popular than others, and everyone comments on them. I also wonder what these topics are, if they are addressed in the same week on both reddit and twitter, and how much of a difference there is if there are separate agendas. As a result, my third research question is, "Can we grasp the weekly agenda and is there a difference between the agendas of the two platforms?"

Contributions of this project follows:

- The creation of a unique small- scale dataset that combines data from online social networks such as Reddit and Twitter.
- The sentiment and similarity analysis of the data collected from social media platforms.

The following is how the paper is constructed in detail: In section 2 I briefly discuss datasets. In section 3 I introduce related work in this area. In section 4 I describe experiments used for analysis of the data. In section 5 I introduce findings of the experiments Finally, in section 6&7 I conclude paper with discussion and conclusion part.

## 2. DATASETS

### *2.1 Reddit Dataset*

Reddit dataset is collected with the help of pipelining mechanism that I did for first project. Dataset has *2023811* posts which makes it about 1.18 GB of data. Table 1 shows the which fields are collected.

**Table 1: Collected fields from the Reddit.**

| Name | Description |
|---|---|
| subreddit | Subreddit that the title belongs |
| title | Title of the post |
| self-text | Text of the post |
| Permalink | URL |

### *2.2 Twitter Dataset*

The Twitter dataset was collected with the sample stream API provided by Twitter. As of right now, Twitter dataset has *1045725* Tweets which makes this dataset 1.11 GB. Table 2 shows which fields are collected during the process

**Table 2 : Collected fields from the Twitter.**

| Name | Type | Description |
|---|---|---|
| tweet.fields | enum(lang, context_annotaions, public_metrics, possibly_sensitive, source) | Enables developer to select which specific tweet fields will deliver |
| expansions | enum(author_id) | Enables developer to get additional data that relate to the originally returned tweets |
| user.fields | enum(public_metrics, verified) | Enables developer to select which specific user fields will deliver in each returned Tweet. |

# 3. RELATED WORK

Relevant related work in this field may be separated into two categories: sentiment analysis and predicting similarities in social networks.

## 3.1 *Predicting similarities in social networks*

Michael Steurer and Christoph Trattner clustered and calculated user's small world characteristics locally [2] by using Jaccard's Coefficient in their project Acquaintance or Partner? Predicting Partnership in Online and Location-Based Social Networks. They conducted their experiments by examining the data they obtained from online social network, Second Life.

On the other hand, Danielle H. Lee, PHD and Peter Brusilovsky, PHD made research about which similarity measure is the most representable for the common interests of multifaceted information among online social connections using the CiteuLike one of the most popular bookmark system.[4]. They examined and compared performance of seven similarity measurement system. To calculate item-based similarity, they used Jaccard Coefficient.

On the other hand, according to Edwin Horlings, PHD standard similarity (or dissimilarity or distance) metrics can be used to compare data points. Because social media users have characteristics (such as age, nationality, and gender), these characteristics might include continuous data (such as age), binary data (such as gender), or categorical data (e.g. nation). The difficulty is to integrate them into a single measure of similarity. If all the data in the measurement is binary and each attribute can only occur once, the Jaccard index can be utilized.[1]

## 3.2 *Sentiment analysis*

Neri, F. at all. states in their paper consumer preferences are influenced by social media through modifying their attitudes and behavior. Monitoring social media activity is a wonderful approach to determine client loyalty and maintain their sentiments about companies or items.[5]

# 4. METHODOLOGY

## 4.1 *Technologies used*

### *4.1.1 Vader sentiment library*

Vader Sentiment library is a rule-based sentiment analysis tool that is specially designed for social media.[3]

### *4.1.2 Matplotlib*

Matplotlib is a Python package that allows you to create static, animated, and interactive visualizations. Matplotlib makes simple things simple and difficult things possible.

### *4.1.3 Csv library*

The csv module provides classes for reading and writing CSV data. It enables programmers to say things like "put this data in the Excel format," or "read data from this Excel-generated file."

### *4.1.4 Re library*

A regular expression describes a collection of strings that match it; the methods in this module allow users to see if a specific string matches a specific regular expression or if a particular regular expression matches a specific string.

### *4.1.5 Word cloud library*

Library for creating word cloud from specific text data.

## 4.2 *Twitter Script*

The script prepared for the analysis of Twitter data consists of several parts. First, the relevant field is retrieved from the database in order to perform sentiment analysis on the tweets, and then it is divided into sentences. Each sentence is sent to the polarity score calculation function of the Vader library. After making function calculations, it calculates negativity, positivity, neutrality, and compound scores. Thus, it is determined whether the text contains a positive, negative, or neutral meaning. Then, if the compound value is greater than *0.05*, it is saved to the database as positive, if it is less than *-0.05*, as negative, and in other cases as neutral.

Then, for word cloud and similarity analysis, alphanumeric characters are cleaned with the help of the text re library and an excel file is created using the csv library from these texts. Finally, the word cloud is created by reading the texts from the excel file.

## 4.3 *Reddit Script*

In the Reddit data measurement and analysis script, sentiment analysis is performed in the same way as in the Twitter script, the only difference is that this analysis is applied on the "title" field for reddit. The process of clearing the title field from non-alphanumeric characters and creating an excel file is like the process done in the Twitter script.

On the other hand, a different process is used to determine the text/link ratio of the reddit self-text area. After the self-text field is retrieved from the database, it is checked whether there is *"https://"* in the self-text field with the help of the search function of the *re* library, and this flag is saved to the database together with the relevant id.

### 4.4 Jaccard Script

The Jaccard coefficient or index, which is defined as the size of the intersection divided by the size of the union of the sample sets, assesses similarity between finite sample sets[w]

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

**Figure 1: Jaccard index formula**

The Jaccard distance that which quantifies dissimilarity between sample sets, is calculated by subtracting the Jaccard coefficient from 1 or, alternatively, dividing the difference between the sizes of the union and the intersection of two sets by the size of the union[6]

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

**Figure 2:Jaccard distance formula**

Jaccard distance is created in four steps in the Jaccard script. First, two sets are created by reading the sentences from the csv file created in the previous stages. It is then calculated and recorded with the help of three different functions that apply the Jaccard formulas.

## 5. RESULTS

### 5.1 Sentiment Analysis

In the first research question, I assumed that social media users were making more negative comments by hiding behind the anonymity provided by social media. That's why I wanted to do sentiment analysis on Reddit titles and tweets. According to the results of the analysis, the ratio of positive tweets to total tweets is *0.4729*. This shows that almost half of the tweets sent have a positive meaning. On the other hand, the ratio of negative tweets to total tweets is *0.2026* and the ratio of negative tweets to total tweets is *0.3243*. Table 3 shows the exact number of positive, negative, and neutral tweets.

Moreover, the ratio of positive Reddit titles to total titles is *0.3404*. The ratio of negative titles to total titles is *0.1489*. The ratio of neutral titles to total titles is *0.5106* Table 4 shows the exact numbers.

**Table 3: Exact numbers of sentiments Tweet analysis.**

| Sentiment | Number of Tweets |
|---|---|
| Positive: | 494585 |
| Negative: | 211965 |
| Neutral: | 339175 |
| Total: | 1045725 |

**Table 4: Exact numbers of sentiments Reddit analysis**

| Sentiment | Number of Reddit Titles |
|---|---|
| Positive: | 688991 |
| Negative: | 301385 |
| Neutral: | 1033435 |
| Total: | 2023811 |

### 5.2 Text/Link Analysis

According to the results of analysis that is made for the second research question, the ratio of self-text field that consists of website link to total posts is *0.9941*. The ratio of self-text field that consists of text to total posts is *0.0058*. Table 5 shows the exact numbers.

As it can be understood from the results, there is a huge difference. Also, it should not be forgotten that the fact that the self-text field consists of a link does not mean that there will be a link that directs the user to another website in that field. Since I started to discover Reddit newly I did not know that at first, I thought it is a link that directs user to another website.

**Table 5: Exact numbers of text/link**

| | Number of self-texts |
|---|---|
| Text: | 2011999 |
| Link: | 11812 |
| Total: | 2023811 |

### 5.3 Similarity Analysis

The third research question was asked in order to understand how similar the agendas of the two platforms were, as the data were collected at the same time interval on both platforms. First, the plan was to create a word cloud from the two platforms separately, each covering the top 200 words, and then measure the similarity with the help of the Jaccard index. However, due to the processing power, it took long time to identify the words and could not be completed. That's why this

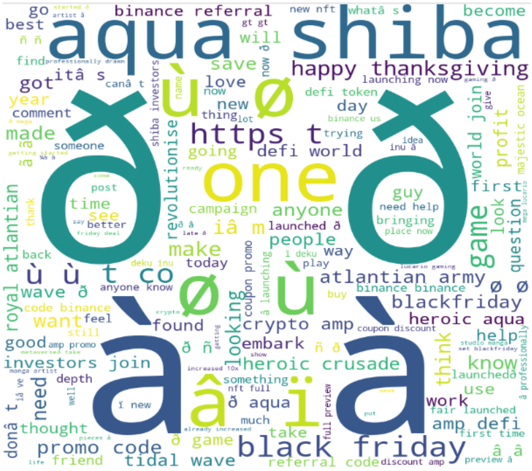

**Figure 3: Reddit title word cloud**

experiment failed. Figure 3 shows Reddit word cloud for nearly %5 of the data before cleaning from regular expressions.

## 6. DISCUSSION

The positive and negative effects of anonymity on the Internet are a controversial issue. The Council of Europe, in its declaration on freedom of communication on the Internet in 2003, states that the wishes of Internet users not to reveal their identities in order to freely share their opinions should be respected.[8]

I prepared my first research question assuming that social media users post more negative comments using anonymity power. But the analysis of the dataset I have shows that positive comments are almost twice negative comments. This means that my initial assumption was not correct, social media is a huge ocean and its positive impact is greater than its negative impact. Moreover, in the future, if sentiment analysis is done by grouping users according to follower count, popular accounts that try to mislead users can be detected, for example conspiracy theorists.

On the other hand, if the self-text/link ratio analysis of reddit is done in the future in order to detect links that lead to other websites apart from, unwanted advertisements can be detected.

Finally, although the analysis of whether the agendas of the two platforms in the same period are similar to each other has not been completed, I think it has an important place in this project. In the future, this analysis can be completed and turned into a commercial social media analysis program. In this way, it is possible to understand easily that what the agenda is or what is the differences of agendas on which platform, between the desired time periods.

## 7. CONCLUSION

In this paper, I conducted investigation for measuring sentiments and similarities on social media data that is collected both Reddit and Twitter. I have shown that the rate of positive comments is higher than the rate of negative comments and the self-text field of Reddit, which consists of only links, is much more than those that consist of text. But I state it again, the fact that it consists of a link does not mean that that link leads to another website and that there is no text in that area.

## 8. REFERENCES

[1] Horlings, Edwin. (2014). Re: What metric can I use to compare the similarity of two social network users?. Retrieved from: https://www.researchgate.net/post/What-metric-can-I-use-to-compare-the-similarity-of-two-social-network-users/545775f5d3df3ee3088b45f1/citation/download.

[2] Steurer, M., & Trattner, C. (2013). Acquaintance or Partner? Predicting Partnership in Online and Location-Based Social Networks

[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] Lee, Danielle H. and Brusilovsky, Peter (2017) How to measure information similarity in online social networks: A case study of Citeulike. Information Sciences, 418-41. pp. 46-60. ISSN 00200255

[5] Neri, F. at all. (2012). Sentiment Analysis on Social Media 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[6] Wikipedia. 2021. Jaccard Index Retrieved Nov, 27 2021 from WikipediA: the Free Encyclopedia. Retrieved from https://www.wikipedia.org/.

[7] Statista. 2021. Facebook - Statistics & Facts Retrieved Nov 27 2021 from https://www.statista.com/topics/751/facebook/#dossierKeyfigures

[8] The Council of Europe .2003. Declaration on freedom of communication on the Internet Retrieved Nov, 27 2021 from https://rm.coe.int/16805dfbd5