# Data Collection System Report

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

## ABSTRACT

Following the rapid developments in information and communication technologies in recent years, devices that can access the internet have now become a part of daily life, used by everyone. The increase in the number of mobile devices has resulted in a special phenomenon that we call social media, which somehow surrounds human life in recent years, becoming the center of daily life, especially the young population. YouTube, Instagram, Facebook, WhatsApp, Reddit and Twitter are the most preferred social media networks.

We're all steadily becoming social media addicts today. Social media has become ubiquitous, whether it's scrolling through our endless Twitter or Instagram feeds on the vehicle or posting the well thought out phrase. In this light, it is reasonable to conclude that the focus of social media and networking is follow-up and sharing. These two notions also serve as the primary data sources for social media analytics, a specialized branch of data analysis.

## Keywords

Social media data, data collection, Twitter API, Reddit API, python, requests, pipelining.

## 1. INTRODUCTION

One of the most crucial aspects of human life is socialization. In contrast to face-to-face communication, social media helps people socialize by giving them a sense of privacy, thanks to its anonymity. Individuals become much more open as a result of their anonymity.[7].

Social media contents are evaluated in the category of user-generated content and are very valuable for businesses, government institutions and individuals. For example, Businesses can perform many activities such as getting an idea about the product and/or brand image with the data they collect through social media, developing a marketing strategy by better understanding the customer, and updating their production processes by identifying potential problems in their products. Or, it can be said that users make purchases based on e-commerce comments, and most of the time, the number of positive comments on the product even compensates for a certain increase in the price of the product.[8 ]

In the light of this information, we can say that we can use social media data in many areas by making social media analysis, that is, by revealing the hidden value. In this study, we will see how a pipelining mechanism can be made to collect social media data, what kind of difficulties can be encountered, which tools can be used etc.

## 2. IMPLEMENTATION

### 2.1 Sources

Two social media platforms were chosen to be used in this project: Twitter and Reddit. Twitter is a microblog and social network platform that allows users post 240 characters limited tweets. [3]. Reddit is a social news community which includes content rating, discussions etc. It is like a huge forum; everything has their own fan page. Ideal for finding own tribe. Users can post their contents as a picture, gif, links, and text.

### 2.2 API's

The use of API is required to collect data from these sources. API is an interface that enables the capabilities of an application, service and/or platform to be accessed externally within the allowed limitations.[5]. Twitter and Reddit both provide API's for accessing to public data of their platform. To be able to use these both API's, authentication is required. All new developers must apply for a developer account to access the Twitter developer platform and Reddit developer platform. Once approved, developer can begin to use new Twitter API v2, or Reddit API. For Reddit, client id, secret key, username, password, and user agent is required. Reddit provides these after applied for developer account. To be able to connect to endpoints of the twitter API, bearer token is required. Also, Twitter provides this information after developer account application is approved.

#### *2.2.1 Twitter's Sampled Stream*

Twitter provides sample stream which delivers approximately 1% of the tweets in real time.[6] In this study, this feature is used in the twitter script. Also, the query parameters used are given in the table below.

**Table 1: Used query parameters in 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. |

### *2.2.1 Reddit's Listing API*

Reddit does not provide streaming API but offers a nice json API. In this project, to grab the most recent posts, reddit's listing API that delivers all new posts with little modification is used. Also, fields used are given in the table below.

**Table 2: Used fields in Reddit.**

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

## 2.3. Technologies used

Python is used in this project as a programming language. Also, to be able to connect to endpoints and get a readable result, some libraries are needed.

### *2.3.1 Requests*

Requests is a python library for HTTP and helps user to send and receive HTTP requests easily.

### *2.3.1 Base36*

Base36 is a library that converts strings that is base 36 to decimal for mathematical operations. Base 36 string consists of ten digits and 26 letters.

### *2.3.1 Bson and Json*

To be able to parse returned data, Bson and Json libraries are required.

### *2.3.1 PyMongo*

Database is required to store the data that retrieved from both sources. Since NoSQL is easy to use, works with json data and python smoothly. Mongo DB is selected to store retrieved information.

## 2.4 Methods

### *2.4.1 Twitter Script*

After authentication information downloaded from the Twitter's developer account. 4 functions were created to handle the process. First function is responsible for creating the URL to connect with requests library. This URL consists of the query parameters that is listed in the section 2.2.1. Table 1.

Second function is responsible for bearer auth. It defines headers with the bearer token that is provided by the Twitter. This function will be called in the connect to endpoint function to provide credentials for the HTTP request.

Third function is the main part. It gets URL and headers, then connects to the URL with the help of the request library. If the response is successful, it prints the response code 200 on the screen and starts getting the data. Then, with the help of json library, json_response file is created to be able to get Tweet object easily. If the language of the Tweet is English, lang parameter that is given in the section 2.2.1, this data is inserted to the MongoDB collection after parsed with the help of bson library which is the fourth function of the script.

### *2.4.1 Reddit Script*

After authentication information downloaded from the Reddit developer account, connected to the url "**https://oauth.reddit.com/r/all/new**" with the parameter "?limit=1" using requests library. Unlike the Twitter, Reddit does not provide streaming services, so stream-like solution needs to be done. In the reddit API, every json response before and after fields that keeps track of post's id so in our approach queries needs to arrange sequentially according to the ids.[1] This is why first connection made by "?limit=1" and the id was taken.

On the other hand, IDs are in base36, so they are needed to convert to decimal to be able to make arithmetic operations. With the help of the base36 library, IDs are converted to decimal and subtracted by 100 then converted to base36 again. In this way, results are obtained as a batch of 100. Then, by making multiple requests, these ids were sent as parameters, respectively to the URL "https://oauth.reddit.com/r/all/new?limit=100&after=t3_ "**.**After parsing the relevant fields from the returned json response, they were saved in the database. Relevant fields can bee seen in the section 2.2.1 Table 2.

## 3. CHALLENGES

Since I started working on the Twitter script for the first time and the documentation of the Twitter API was very well done, I did not encounter any problems. However, I stated that I would use tweet, media, and user fields in the proposal. Then I chose not to use the media field because I had no idea how to analyze it. I wanted to use the public metrics of the user field instead of the media field, so I added the expansion field to my query.

The real difficulties arose when I started working with the Reddit API. The first problem that arose was the authentication problem. Reddit somehow couldn't authenticate my User-agent when I used the request library but could authenticate when I used the praw library. On the other hand, I couldn't use the crawling tool as it is forbidden to use it. I solved this problem by deleting and recreating the reddit developer credits a few times. Also, Reddit's API documentation was more difficult to understand and interpret because it was complicated and not beautiful, so I used other sources while creating the Reddit script. Since the examples in the sources I used were not in Python language, it was not easy to interpret and cost me some time.

## 4. DISCUSSION & CONCLUSION

Anyone, depending on where technology has progressed, may be able to access social media. In the digital environment, factors such as secrecy and anonymity make it easier to satisfy the demand for sociability. On the other hand, social media makes companies earn significant amount of money because of the data that shared by users. To be able to collect this data, in this paper, I presented and implemented social media data pipelining mechanism for the two of the social media platforms, Reddit and Twitter. By the end of the semester, I am expecting to collect 20-25 million reddit posts and 170-180 million tweets. For the collected data to gain meaning, the measurement and analysis system will be implemented as a next project.

## 5. REFERENCES

[1] Lewis Van Winkle, 2018. How F5Bot Slurps All of Reddit.(July 2018).Retrieved November 5, 2021, from https://intoli.com/blog/f5bot/

[2] Reddit, 2020, Reddit :Dive into Anything. Retrieved from https://www.reddit.com/r/blog/comments/k967mm/reddit_in_2020/

[3] Wikipedia. 2017. WikipediA: the Free Encyclopedia. Retrieved from https://www.wikipedia.org/.

[4] David Sayce,2020. The Number of tweets per day in 2020. Retrieved October 3, 2021, from https://www.dsayce.com/social-media/tweets-day

[5] Perry Eising,2017. What exactly is an API. Retrieved November,5, 2021, from https://medium.com/@perrysetgo/what-exactly-is-an-api-69f36968a41f

[6] Twitter Developer Platform. Sampled streams, Retrieved November 5, 2021, from https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction

[7] Varma, V., Kurisinkel, L. J., & Radhakrishnan, P. (2017). Social Media Summarization. InCambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds), A Practical Guide To SentimentAnalysis (pp. 135-153). Springer, Cham.

[8] Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. Information fusion, 36, 10-25