**Marios Petrov  
 Computer Science/Neuroscience  
 New College of Florida  
Sarasota, FL**[**marios.petrov23@ncf.edu**](mailto:marios.petrov23@ncf.edu)

**Shawn Nash  
 Computer Science  
 New College of Florida  
 Sarasota, FL**

[**shawn.nash24@ncf.edu**](mailto:shawn.nash24@ncf.edu)

**ABSTRACT**

During this independent study project (ISP), we set out to examine the efficacy of data-mining/scraping techniques on three different social media platforms. Our findings revealed that the most data rich platform, Facebook, is the most difficult and convoluted platform for researchers to scrape and gather data from. On the other hand, Twitter and Reddit were substantially more   
“scrapper-friendly” and enabled researchers to utilize the plethora of meta data available. This polarity between platform policies on the gathering of meta-data, can potentially have long lasting repercussions on the validity of social media NLP research. A standardized tool/method for researchers to access metadata on Facebook would effectively bridge this existing polarity.

**INTRODUCTION AND CONCEPTS**

***Social Media Data Mining***

Social Media Data Mining is the process of analyzing and extracting valuable information from social media platforms such as Facebook, Reddit, Twitter, Instagram, and LinkedIn. This method involves using specialized software tools to gather, process, and evaluate social media data in order to uncover meaningful information.

***Social Media Data Scraping and Extraction***

Social Media Data Scraping or Extraction refers to the automated process of collecting and gathering data from social media platforms. Data scraping is used to extract large volumes of data, such as posts, comments, and user profiles, and to analyze them for specific insights or patterns. Social Media Data Scraping can be done using specialized software tools, such as web crawlers, or through programming languages like Python.

***The Valuable Data of Social Media***

Social media sites are valuable data resources because they offer a vast and diverse array of information that can be used to gain insights into consumer behavior, market trends, and public opinion. These platforms provide real-time access to millions of users around the world, making them ideal for studying social phenomena or collecting data on human behavior. Moreover, social media platforms provide data that is rich in context and can be used to develop a more complete understanding of the world we live in. As a result, social media data has become an increasingly valuable resource for researchers, marketers, and businesses alike.

***Social Media Datasets, Deep Learning, and Data-Oriented Programming***

Social Media Mining involves the collection and analysis of large datasets from social media platforms like Facebook, Reddit, and Twitter. These datasets contain a wealth of information, including user profiles, posts, and comments. Deep learning neural networks, such as the Valence Aware Dictionary and sEntiment Reasoner (VADER), are often used in Social Media Mining to analyze the sentiment of social media content. These neural networks use natural language processing (NLP) techniques to determine the emotional tone of a particular text or post. Data-oriented programming and design is integral to Social Media Mining, as it involves developing software applications that can effectively manage and analyze large datasets. Data-oriented programming focuses on designing software programs that are optimized for data processing, storage, and retrieval, ensuring that data analysis can be done quickly and efficiently.

***Approaching Social Media Web Scraping***

When approaching Social Media Web Scraping, it is essential to follow the best practices to ensure that the research is conducted ethically and in compliance with the social media platform's terms of use; this can vary for differing social media sites. Before beginning any data collection, the researcher should be clear on the research objectives, target social media platforms, and the types of data to be collected. It is suggested to have a well-defined methodology for data collection and processing, including any necessary data cleaning or preprocessing. Additionally, it is important to be mindful of potential ethical concerns, such as data privacy and consent. Finally, it is important to consider any legal issues related to the use of web scraping tools and data ownership rights, and take necessary precautions to ensure compliance with relevant laws and regulations.

**RELATED WORKS**

Social media mining is a rapidly evolving field that has drawn inspiration from various past literature and research in the fields of data science, machine learning, and computational linguistics. With such inspirations to the field, there are a few notable works that include motivation to the research in this paper.

The first work, "Social Media Web Scraping using Social Media Developers API and Regex" by Lusiana Citra Dewi, Meiliana, and Alvin Chandra3, focuses on web scraping techniques for social media data using the social media developer API and regular expressions (regex). This work is highly relevant to the scraping of Facebook, Reddit, and Twitter as it highlights the importance of using APIs and regex to extract and preprocess social media data.

The second work, "What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data" by Moreno Mancosu and Federico Vegetti4, proposes a tool to collect public Facebook data while also discussing the ethical implications of web scraping. This work is highly relevant to the scraping of Facebook as it emphasizes the importance of ethical considerations when collecting and analyzing social media data.

The third work, "Reducing the Noise From Scraping Social Media Content: Some Evidence-Based Recommendations" by Filip Lievens and Chad H. Van Iddekinge5, provides evidence-based recommendations for reducing noise in social media data. This work is highly relevant to the scraping of Reddit and Twitter as it provides guidelines for improving the quality of data collected through web scraping.

Overall, the work of scraping Facebook, Reddit, and Twitter draws inspiration from and builds upon past literature in social media mining by highlighting the importance of using APIs, regex, ethical considerations, and data quality. By considering and implementing these factors, researchers can ensure the accuracy and validity of the data collected through web scraping, which can provide valuable insights into various fields, such as business, politics, and public health.

**PLATFORMS EXPLORED**

***Facebook***

Facebook's API allows us to access a vast amount of data on public pages and groups, including posts, comments, likes, and shares. However, access to user data is limited due to privacy concerns, and scraping data from private groups or personal profiles is not allowed. Additionally, Facebook's API is subject to rate limits and quotas, which can restrict the amount of data that can be collected in a given time frame. Despite these limitations, Facebook's API provides valuable data on user behavior and social dynamics, making it a valuable resource for web scraping and data collection.

Regarding metadata, Facebook provides quite the structure of metadata that can be collected through web scraping. This metadata includes information about user profiles, such as demographic data, interests, and social connections, as well as data on the structure and content of pages and groups. However, it is important to note that collecting and analyzing metadata from Facebook must be done with caution and in compliance with ethical and legal guidelines to protect user privacy and ensure the integrity of the research.

Accessing Facebook's data through web scraping is challenging due to several factors, including the company's strict policies on data collection and privacy concerns. Given the challenges and limitations of scraping data from Facebook, we decided that it may not be the best option for data collection at this time. As such, we made the decision to explore other sources of social media data or consider alternative methods of data collection to collect our datasets and to achieve our research goals.

***Reddit***

When considering data mining from Reddit, it is important to weigh the pros and cons of using its API. One significant advantage of using Reddit's API is its ease of use and the availability of a wide variety of data types, such as posts, comments, and user information. Additionally, Reddit's API is designed to facilitate data scraping and does not have as many restrictions as other social media platforms, making it relatively easy to collect a large amount of data. However, some potential drawbacks include the unstructured nature of the data and the potential for biased data due to the user-based nature of the platform. Due to the lack of structure on the data we were scraping, we had to limit the dataset to just the main posts from subreddits, due to the increased time it would take to create a larger dataset including the comments from each scraped post.

In terms of metadata, Reddit provides a generous set of data that can be collected through web scraping. This metadata includes information about user profiles, such as the date they joined, the subreddits they follow, and the karma they have earned. Additionally, Reddit's structure allows for the collection of data on individual subreddits, including the number of subscribers, the types of posts, and the frequency of posting. This metadata can be used for various research purposes, such as social network analysis, content analysis, and sentiment analysis, among others. Overall, Reddit's API and metadata provide a wealth of data that can be used to gain insights into user behavior and social dynamics on one of the largest social media platforms in the world.

The ease of access to Reddit's API and their well-documented Python library proved to be advantageous in our scraping and dataset creation process. The API allowed us to retrieve a large amount of data in a structured and organized manner, which was critical to the accuracy and completeness of our dataset. Additionally, the use of Python allowed for efficient and effective scraping, as well as the ability to clean and process the collected data. The well-documented Python library also made it easy to navigate and utilize the API, making it an ideal choice for us looking to collect data from social media sites. All of this stayed true as long as we stuck to scraping the data from just the posts; having us excluding the comments from our dataset.   
  
One of the challenges when scraping Reddit's API that made us look into and consider using Twitter’s API was the potential for certain subreddits’ top posts to be primarily image-based posts rather than text-based ones. This makes it difficult to collect data for sentiment analysis or text-based research.

***Twitter***

Twitter's API provides a wealth of data that can be leveraged for social media research. One of the main advantages of using Twitter's API is the ability to collect real-time data on a wide range of topics. The platform's popularity and global reach mean that it is often used to share breaking news and opinions on current events, making it a valuable resource for researchers interested in tracking trends and sentiment around specific topics. Another benefit of using Twitter's API is the ability to collect a wide range of metadata, including information on user profiles, tweets, retweets, and followers. This metadata can provide valuable insights into user behavior, engagement, and sentiment. Additionally, Twitter's API provides access to historical data, making it possible to analyze trends and patterns over time. One of the main drawbacks of using Twitter's API, however, is the potential for bias in the data. Because Twitter users are not representative of the general population, it is important to carefully consider the sample of users and topics being studied.

In terms of metadata that can be collected from Twitter, there are several key pieces of information that can be useful for research purposes. These include user profile information, such as username, location, bio, and number of followers. Tweet metadata can include the content of the tweet, the date and time it was posted, the number of likes and retweets, and the user who posted it. Additionally, Twitter's API provides access to metadata on trending topics, hashtags, and mentions, which can be useful for tracking and analyzing social media trends.

**DATASETS AND METHODOLOGY**

Our research focus was more exploratory and less analytical in nature. As stated, we aimed at assessing how each social media platform enables or disables social media data scraping. However, in our scraping efforts we constructed the Nutrition subreddit dataset and the HowToBuildAPC subreddit dataset.

***Scraping Methodology***

When looking to create our dataset by scraping posts from Facebook, Reddit, and Twitter using their respective APIs, we had to consider the different challenges and the different approaches required for each social media site.

***Facebook***

Facebook looked promising at first with the vast amount of data on the site alone. Facebook's GraphAPI provides a relatively easy way to scrape posts from public pages and groups, however access to user data is limited due to privacy concerns which makes it very difficult to scrape clean data from Facebook. Facebook’s GraphAPI also uses a built-in browser tool and provides very little information, as well as lacking necessary Python libraries and documentation to use such tools and to create our datasets.

***Reddit***

Reddit, on the other hand, made scraping posts and comments to make a dataset much easier using their API. Reddit's API provides access to a vast amount of data on user-generated content and social interactions on the platform, including posts, comments, votes, and user profiles. This data can be leveraged for various research purposes, such as sentiment analysis, topic modeling, and social network analysis, among others. The API comes with a Reddit Python library as well as full documentation, allowing for targeted and efficient scraping of data, enabling us to collect large amounts of data with minimal effort.

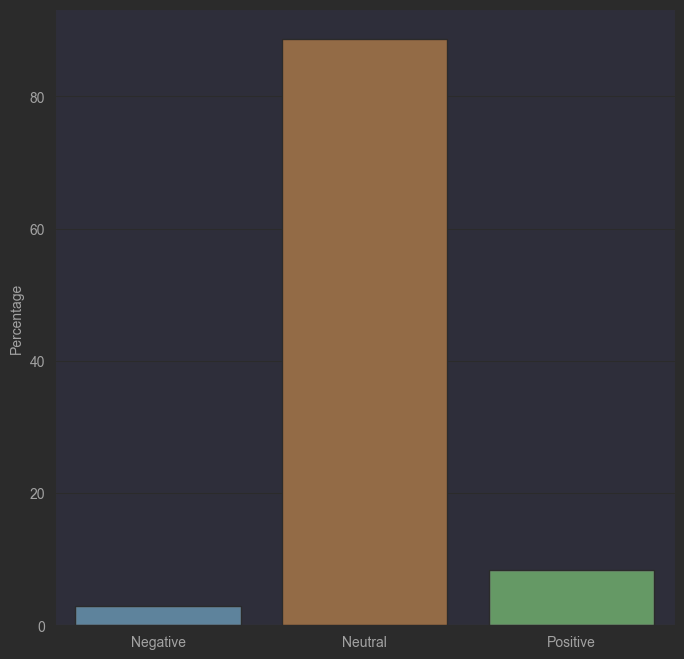
***Twitter***Twitter's API provides access to a vast amount of real-time and historical data on user-generated content, social interactions, and network structures on the platform. However, Twitter’s API is fairly similar to Reddit’s API in the sense that the data provided by Twitter's API can be useful for conducting research on various topics, such as analyzing sentiment, identifying opinions, detecting events, and managing crises. With the API, akin to Reddit’s API, it is also possible to scrape data in a focused and efficient manner, which allows for the collection of large amounts of data with minimal effort.

**DATA ANALYSIS**

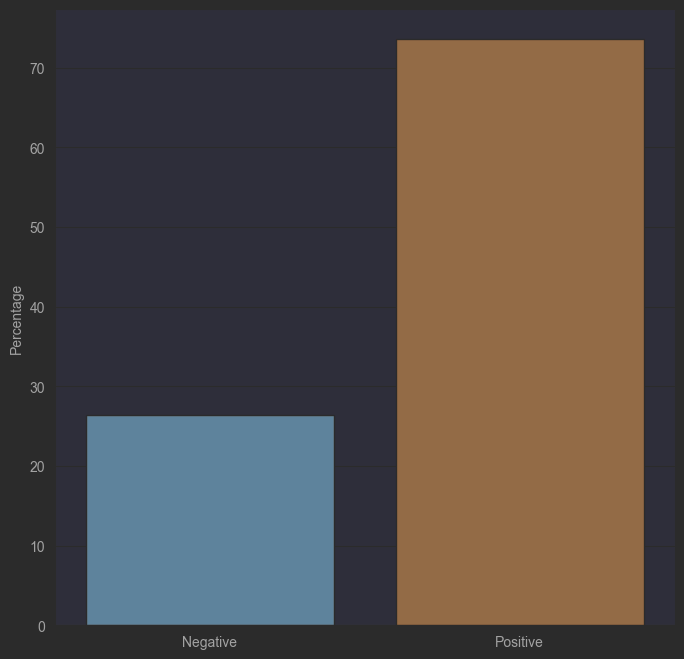
Both datasets underwent preprocessing, which consisted of: tokenization, case-standardization (lower), stopword removal, url removal, and emoji removal. Stemming or lemmatization wasn’t utilized due to some peculiar interactions with terms present in the dataset. Furthermore, we performed sentiment analysis on the datasets and generated polarity scores for each word using VADER.

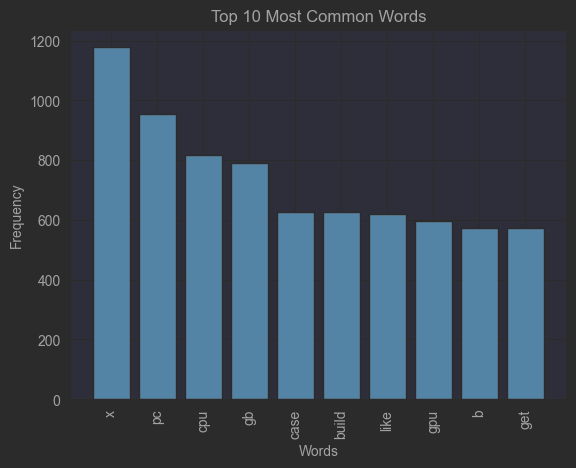
***HowToBuildAPC Dataset Analysis***

The HowToBuildAPC dataset initially demonstrated a severe skew towards neutral words, in terms of polarity scores; however, after removing neutral terms, the dataset then represents a skew towards positive words. After reviewing the frequency distribution for the most common words, it is uncertain as to whether this skew in the data is due to the actual structure of the dataset itself, or as a result of potential preprocessing side effects.



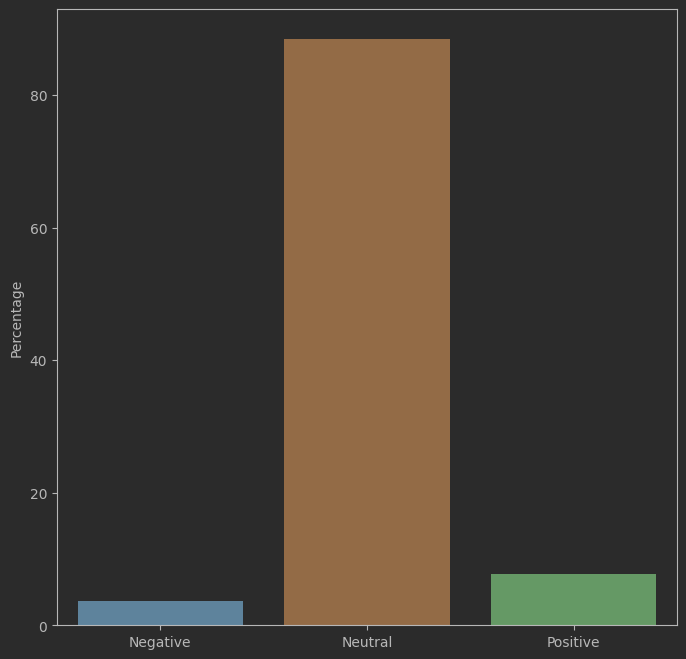
**Figure 1:** Bar graph displaying the polarity score distribution of the HowToBuildAPC Subreddit dataset (negative = blue, positive = green, neutral = brown). Polarity scores were generated by VADER, a pre-trained sentiment analysis model. As expected, the dataset is immensely skewed neutral, thus further processing is advised to fully take advantage of this data.

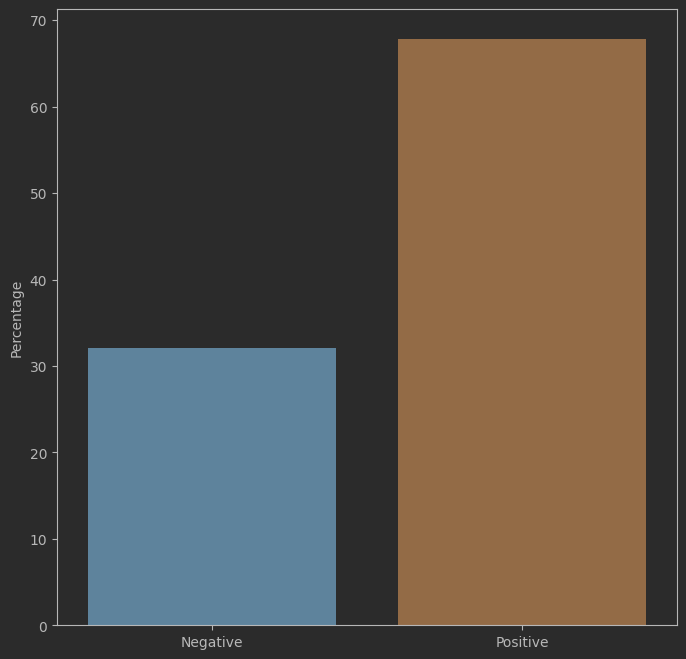
**Figure 2:** Bar graph displaying the polarity score distribution of each word within the preprocessed HowToBuildAPC Subreddit dataset (negative = blue, positive = brown). Polarity scores were generated by VADER, a pre-trained sentiment analysis model. After removing neutral words from the dataset, there still remains an immense skew towards positive phrases.

**Figure 3:** Distribution displaying the top 10 most common words within the HowToBuildAPC Subreddit dataset (x-axis: words, y-axis: frequency).

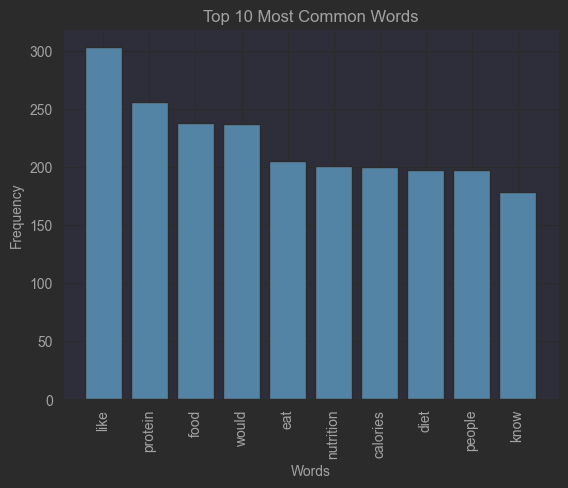
***Nutrition Dataset Analysis***

Similarly to the HowToBuildAPC dataset, the “Nutrition” dataset also exhibits a large skew in favor of positive terms, when neutral terms are accounted for. Unlike the other dataset, there were no single letter words/tokens present in the most frequent words distribution. This leads us to speculate that the terms “x” and “b” present in the previous dataset, were by-products of our preprocessing methods.

****Figure 4:** Bar graph displaying the polarity score distribution of each word within the preprocessed Nutrition Subreddit dataset (negative = blue, positive = green, neutral = brown). Polarity scores were generated by VADER, a pre-trained sentiment analysis model. As expected, the dataset is immensely skewed neutral, thus further processing is advised to fully take advantage of this data.

******

**Figure 5:** Bar graph displaying the polarity score distribution of each word within the preprocessed Nutrition Subreddit dataset (negative = blue, positive = brown). Polarity scores were generated by VADER, a pre-trained sentiment analysis model. After removing neutral words from the dataset, there still remains an immense skew towards positive phrases.

****

**Figure 6:** Distribution displaying the top 10 most common words within the Nutrition Subreddit dataset (x-axis: words, y-axis: frequency).

**DISCUSSION AND FUTURE WORK**

Our simple and relatively small Reddit datasets display what an invaluable resource social media platforms can be for researchers. These platforms are overflowing with metadata that, without properly developed tools, cannot be accessed efficiently, as is the case with Facebook.

In future endeavors, we seek to explore more platforms, and construct a visual hierarchy to represent scrapper-friendliness for each respective platform. An easily accessible resource of this nature would equip researchers with a tool, to make more educated and reasonable choices as to what platform to mine data from for whatever the project may be.

Furthermore, developing more sophisticated and focused data scraping parameters within our scripts would further accelerate the utility of future dataset construction. There is also great potential in conducting individual news-feed based analysis, in an effort to reverse-engineer and improve social media news-feed user algorithms.

**REFERENCES**

**[1]** Huan Liu, Fred Morstatter, Jiliang Tang, and Reza Zafarani. 2016. The good, the bad, and the ugly: Uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics* 1, 3-4 (2016), 137–143. DOI:<http://dx.doi.org/10.1007/s41060-016-0023-0>

**[2]** Pritam Gundecha and Huan Liu. 2012. Mining Social Media: A brief introduction. *2012 TutORials in Operations Research* (2012), 1–17. DOI:<http://dx.doi.org/10.1287/educ.1120.0105>

**[3]** Lusiana Citra Dewi, Meiliana, and Alvin Chandra. 2019. Social media web scraping using social media developers API and regex. *Procedia Computer Science* 157 (September 2019), 444–449. DOI:<http://dx.doi.org/10.1016/j.procs.2019.08.237>

**[4]** Moreno Mancosu and Federico Vegetti. 2020. What you can scrape and what is right to scrape: A proposal for a tool to collect public Facebook data. *Social Media + Society* 6, 3 (2020), 205630512094070. DOI:<http://dx.doi.org/10.1177/2056305120940703>

**[5]** Filip Lievens and Chad H. Van Iddekinge. 2016. Reducing the noise from scraping social media content: Some evidence-based recommendations. *Industrial and Organizational Psychology* 9, 3 (2016), 660–666. DOI:<http://dx.doi.org/10.1017/iop.2016.67>