Analyzing Stock Market Sentiment and Price Movements

Bhargav Patel
bpatel12@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

Harshit Kamlesh Soni
hsoni1@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

Shashank Agrawal
sagrawa8@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

ABSTRACT

In the realm of data science, effective data collection is pivotal. This project's core objective is to establish a robust data collection system to extract valuable insights from two key sources: Reddit and Yahoo, facilitated by Request-HTML. Our project revolves around analyzing sentiment in Reddit discussions pertaining to specific stocks and linking it to their corresponding price movements on Yahoo. Reddit data will be sourced from finance-related subreddits, with flexibility to include more. A Reddit scraper will periodically gather real-time data, including post and comment information. We'll also collect real-time stock market data from Yahoo Finance, including stock prices, volumes, and financial metrics. The Yahoo scraper will extract data from HTML responses. The perpetual nature of financial markets ensures a continuous dataset for analysis. This project lays the foundation for in-depth stock market sentiment and price movement analysis, enhancing our grasp of market dynamics.

KEYWORDS

Sentiment Analysis, Stock Market, Reddit, Yahoo, Request-HTML, Real-Time Data, Perpetual Dataset

1 INTRODUCTION

Building upon our successful establishment of a robust data collection system for extracting valuable insights from Reddit and Yahoo Finance within the context of stock market sentiment and price movements, the next phase of our data science journey is centered on the execution of measurement and analysis experiments. While data collection serves as a foundation, the true essence of data science lies in the discovery of meaningful insights and the application of knowledge derived from data. This is designed to take our data analysis to a deeper level, introducing experiments, and integrating a critical new component for measuring content for mutual relationships and toxicity.

In this project, we set out to accomplish the following pivotal objectives:

- **Designing Measurement Experiments:** We acknowledge that data collection is merely the initial step in the data science process. To unlock the full potential of our dataset, we must design a series of experiments aimed at comprehensively describing the data. These experiments will enable us to gain a deeper understanding of the intricate dynamics within the realm of stock market sentiment and its correlation with price movements.
- **Research Questions:** We will formulate a set of at least three research questions specific to our dataset. While addressing these questions is a component of this project, the primary focus is on developing the methodology and approach for future exploration in Project 3. These research questions will serve as guiding principles for our analysis and experimentation.
- **Measuring Toxicity:** To add a new dimension to our data analysis, we will incorporate real-time measurements of content toxicity. Leveraging the ModerateHatespeech API (https://moderatehatespeech.com), we will assess the level of toxicity in discussions related to stocks on Reddit. This component introduces a critical aspect of content analysis and sentiment evaluation, enhancing our understanding of user-generated content.
- **System Robustness:** It is imperative to ensure the robustness of our data collection system against potential challenges, including the availability of the ModerateHatespeech API. We will address any challenges or limitations that may arise while working with the API, ensuring our system can handle edge cases and outages, thus ensuring uninterrupted data flow and analysis.
- **Real-time Integration:** We will work diligently to achieve real-time integration of toxicity measurements with our data collection system. This step is crucial for ensuring that toxicity scores align seamlessly with the data ingestion process, setting the stage for a more advanced analysis in Project 3.

The project proposal will outline our plan for executing these objectives, taking into consideration potential challenges and the need for additional data. Our aim is to possess a deeper understanding of our dataset and be well-prepared for more advanced analysis in Project 3.

2 DATASETS

This project involved the collection of data from two primary sources, namely Reddit and Yahoo. The following sections outline the details of the data collection process from each of these sources:

2.1 REDDIT

To collect data from Reddit, the following steps were taken:

- Data Retrieval Using Requests: A Python script was developed to interact with Reddit's API using the Requests library. This script made periodic GET requests to access the desired data from specified subreddits, including but not limited to r/wallstreet and r/stocks.

- Periodic Data Collection: The code was designed to run periodically to ensure a continuous stream of real-time data from Reddit. This periodic execution facilitated the collection of up-to-date information.
- Data Structuring: The collected Reddit data was structured into a comprehensive format, such as a dataframe. This structured dataset encompassed essential information from both posts and comments, including post ID, subreddit, post title, post score, comment body, comment score, and timestamps.

2.2 YAHOO

For collecting data from Yahoo, Python and the following techniques were utilized:

- Periodic GET Requests: A Python script was created that periodically sent GET requests to specific pages on Yahoo Finance, targeting the information needed. These requests were necessary to access real-time financial data.
- HTML Data Extraction: Upon receiving the HTML responses from Yahoo Finance, libraries like BeautifulSoup were employed to extract the desired financial data. This included stock prices, trading volumes, and other pertinent metrics.
- Real-time Data Streaming: Similar to the Reddit data collection process, this script ran at regular intervals to maintain real-time data streaming.
- Data Structuring: The collected Yahoo data was structured into a dataset suitable for analysis. This dataset included fields such as date, trending ticker, company name, analyst rating, overall score, and related news.

By implementing these processes and leveraging the specified libraries for both Reddit and Yahoo data collection, the project had a continuous flow of real-time data from these sources. This comprehensive dataset served as the foundation for the in-depth analysis of stock market sentiment and its correlation with price movements, contributing to a deeper understanding of market dynamics and trends.

3 BACKGROUND AND RELATED WORK

3.1 Data Science in Financial Markets

Effective data collection and analysis play a pivotal role in understanding and predicting trends in financial markets. The intersection of data science and finance has become increasingly important, with researchers and practitioners seeking innovative ways to leverage data for market insights. The ability to analyze sentiment in online discussions, particularly in platforms like Reddit, has garnered attention as it offers a unique perspective on public sentiment regarding stocks.

3.2 Sentiment Analysis in Financial Markets

The assessment of sentiment in financial markets has been a subject of extensive research. Numerous studies have explored the correlation between social media sentiment and stock price movements. Researchers have employed various natural language processing (NLP) techniques to extract sentiment from textual data, providing valuable insights into the impact of public opinion on financial markets.

3.3 Integration of Social Media and Financial Data

The integration of social media data, such as discussions on platforms like Reddit, with traditional financial data sources is an emerging area of interest. This integration aims to provide a holistic view of market dynamics by combining quantitative financial data with qualitative information derived from online discussions. Understanding how social media trends align with or deviate from financial indicators contributes to a comprehensive understanding of market behavior.

3.4 Existing Tools and Technologies

Request-HTML, a Python library for web scraping, serves as a key tool in this project for extracting data from both Reddit and Yahoo. The utilization of web scraping techniques aligns with the need for real-time and continuous data collection, crucial for capturing dynamic market changes. Additionally, the periodic data collection intervals, both for Yahoo and Reddit, are designed to ensure a steady stream of updated information for analysis.

3.5 Relevance to the Current Project

This project addresses the gaps in existing research by specifically focusing on the synchronicity between Reddit discussions and stock price movements sourced from Yahoo Finance. While previous studies have explored sentiment analysis in financial markets, the direct linkage between trending stocks and related discussions on Reddit is a novel approach. The methodology employed in this project draws inspiration from existing work but tailors it to the unique context of analyzing the interplay between Reddit and Yahoo data.

Understanding the background and related work in this domain sets the stage for the subsequent sections, where we delve into the specifics of our data collection, integration, and analysis methods. The synthesis of insights from both financial data and online discussions aims to provide a comprehensive understanding of the dynamics between social media sentiment and stock market trends.

4 METHODOLOGY FOR DATA ANALYSIS

In this project, we aim to analyze the synchronicity, or lack thereof, between trending stocks obtained from Yahoo and trending posts related to those stocks on Reddit. The primary objective is to determine which trending stocks from Yahoo are in sync or out of

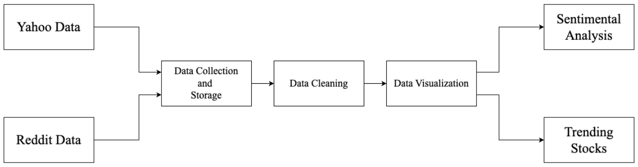


Figure 1: Flowchart

sync with trending news and related comments on Reddit. This analysis will provide insights into how the two platforms reflect and influence each other in the context of stock market trends.

4.1 Data Collection and Storage

4.1.1 YAHOO DATA COLLECTION.

- Continued data collection from Yahoo focused on trending stocks and related news articles. The collected data included information such as stock symbols, company names, analyst ratings, overall scores, and related news articles.
- The collected data was stored in a structured dataset suitable for analysis. This dataset included fields like stock symbol, company name, analyst rating, overall score, and related news.
- Data collection from Yahoo occurred at regular 5-hour intervals to maintain real-time data streaming.

4.1.2 REDDIT DATA COLLECTION.

- Collected data from Reddit, specifically focusing on trending posts related to stocks. This involved scraping data from finance-related subreddits (S1, S2, S3, S4, S5).
- For each subreddit, the 5 latest trending posts were collected. Additionally, comments related to each of the 5 trending posts per subreddit were gathered.
- The Reddit data collected was structured into a comprehensive format, such as a dataframe. This structured dataset included information like post title, post score, comment body, comment score, and timestamps.
- Data collection from Reddit was scheduled at regular 10-hour intervals to ensure a continuous flow of real-time data.

4.2 Data Integration and Synchronization Analysis

In the assessment of synchronicity between Yahoo's trending stocks and Reddit's trending posts and related comments, the following steps were undertaken:

4.2.1 DATA INTEGRATION.

- The Yahoo data and Reddit data were merged using common identifiers like stock symbols and names.
- This integration resulted in a unified dataset that combined information from both sources, aligning trending stocks and Yahoo's news with their corresponding Reddit posts and comments.

4.2.2 SYNCHRONICITY ANALYSIS.

- A metric or scoring system was defined to assess the synchronicity between Yahoo trending stocks and Reddit trending posts.
- The scoring system considered factors such as the frequency of mentions, sentiment analysis of Reddit comments, and alignment of post titles with stock symbols.
- Using this metric, trending stocks were categorized into three groups: in sync, partially in sync, and out of sync with Reddit trends.

5 DESCRIPTIVE ANALYSIS

5.1 Yahoo Data

In this project, we extensively utilized Yahoo data, focusing on stocks' posts identified by their tickers and their associated news articles. Each post corresponds to an individual stock and is linked to relevant news pieces. This integration allowed us to gain comprehensive insights into the dynamics of stock-related discussions on Yahoo.

- **Stock Information:** We stored crucial information in our database, including the stock's ticker name and its corresponding name. This enabled us to uniquely identify and categorize each stock for further analysis.
- **Yahoo News:** Our dataset encompasses a collection of Yahoo news articles related to specific stocks. These articles serve as valuable sources of information, contributing to a more comprehensive understanding of market sentiment and trends.
- **Yahoo Behavior Analysis:** For a nuanced perspective, each Yahoo post's behavior was analyzed and categorized as either negative, positive, or neutral. This sentiment analysis provided a deeper insight into the overall sentiment surrounding each stock on the Yahoo platform.

5.2 Reddit Data

Our project also extensively incorporated Reddit data, focusing on stocks' posts and the ensuing discussions in the form of comments. Unlike Yahoo, each Reddit post can generate multiple comments, allowing for a more dynamic and interactive analysis.

- **Reddit Comments:** We stored both the posts and their associated comments in our database. This wealth of data

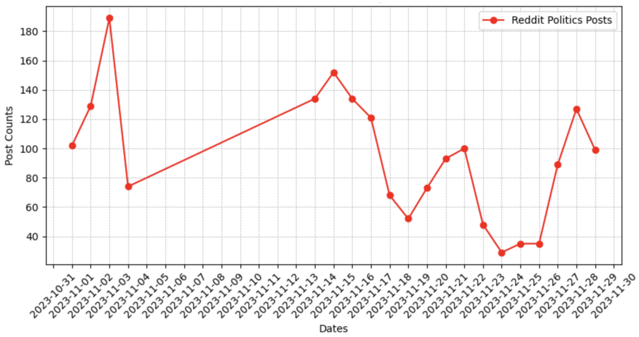


Figure 2: Reddit Politics Posts Activity Over Time

enables us to delve into the community discussions surrounding specific stocks on Reddit, capturing diverse opinions and insights.

- **Reddit Behavior Analysis:** Similar to Yahoo, we conducted sentiment analysis on Reddit comments, categorizing each comment's behavior as negative, positive, or neutral. This step was crucial in understanding the sentiment prevailing within the Reddit community regarding different stocks.

5.3 Data Integration and Confidence Analysis

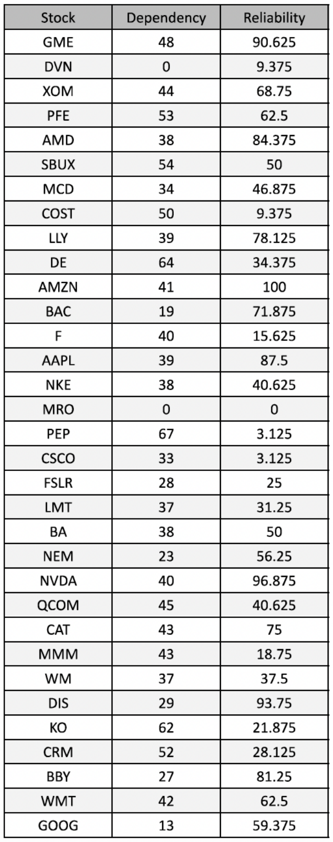
To ensure the reliability of our insights, we created the-`yahoo_reddit_collection` table, consolidating data from both Yahoo and Reddit. This table includes essential fields such as Ticker, Company Name, Yahoo Headline, Yahoo Sentiment, Yahoo Date, Reddit Text, Reddit Sentiment, Reddit Date, and Insertion Date.

- **In-Sync Analysis:** We implemented a comprehensive analysis to determine the synchronization of data between Yahoo and Reddit. If both data sources showed positive sentiment dependencies, we concluded that the data was in sync. Conversely, inconsistencies in sentiment across platforms indicated a lack of synchronization (see Table 1).

- **Confidence Rate:** We introduced a confidence rate as a metric to quantify the reliability of our in-sync analysis. This rate reflects the level of confidence we have in the alignment of sentiment between Yahoo and Reddit data. A higher confidence rate signifies a stronger correlation and reliability in our findings.

5.4 Comparative Analysis

Beyond individual platform analyses, we conducted a comparative study to identify patterns and divergences between Yahoo and Reddit data. This comparative lens provided a nuanced understanding of how sentiments differ or converge across platforms.



Stock	Dependency	Reliability
GME	48	90.625
DVN	0	9.375
XOM	44	68.75
PFE	53	62.5
AMD	38	84.375
SBUX	54	50
MCD	34	46.875
COST	50	9.375
LLY	39	78.125
DE	64	34.375
AMZN	41	100
BAC	19	71.875
F	40	15.625
AAPL	39	87.5
NKE	38	40.625
MRO	0	0
PEP	67	3.125
CSCO	33	3.125
FSLR	28	25
LMT	37	31.25
BA	38	50
NEM	23	56.25
NVDA	40	96.875
QCOM	45	40.625
CAT	43	75
MMM	43	18.75
WM	37	37.5
DIS	29	93.75
KO	62	21.875
CRM	52	28.125
BBY	27	81.25
WMT	42	62.5
GOOG	13	59.375

Table 1: Trading stocks dependency on Reddit Sentiments

5.5 Visualization

To enhance the interpretability of our findings, we incorporated visualizations such as charts and graphs representing sentiment distributions, trends over time, and comparative analyses. These visual aids offer a clearer presentation of the complex data relationships.

5.6 Sync Data Analysis

In our Sync Data Analysis, we focused on the integration of Yahoo and Reddit data, specifically in terms of sentiment synchronization. The `Ticker_Sync_Data` table was generated from the analysis of the `yahoo_reddit_collection` table, with the following key metrics:

- **Ticker:** The stock's ticker symbol.
- **Company Name:** The name of the company associated with the stock.
- **Number of InSync:** The count of instances where Yahoo and Reddit sentiments match (positive, negative, or neutral).
- **Number of OutSync:** The count of instances where Yahoo and Reddit sentiments differ.
- **Dependency:** The percentage of in-sync instances calculated as

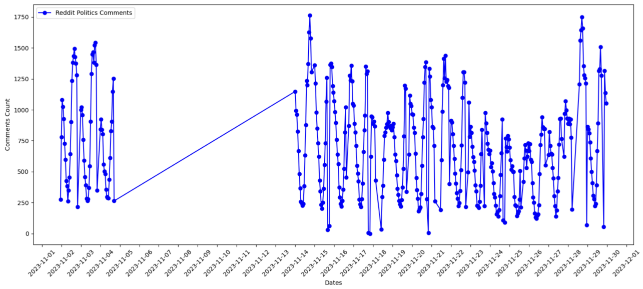


Figure 3: Hourly Reddit Politics Comments Activity

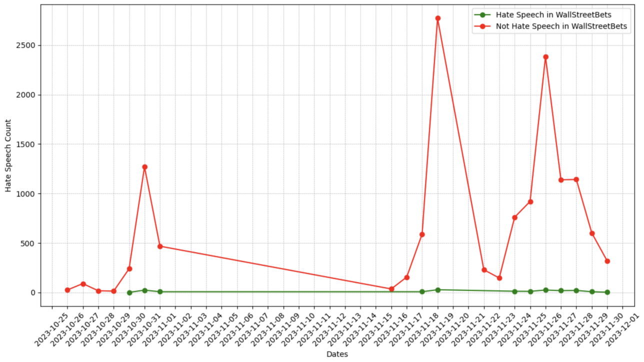


Figure 4: ModerateHatespeech Activity Over Time in WallStreetBets

$$\text{Dependency} = \frac{\text{Number of InSync}}{\text{Number of InSync+Number of OutSync}} \times 100$$

- **Reliability:** The reliability ratio calculated as

$$\text{Reliability} = \frac{\text{Number of InSync+Number of OutSync}}{\max(\text{Number of InSync,Number of OutSync})} \times 100$$

- **Insertion Date:** The date when the analysis was performed.

This comprehensive analysis provides valuable insights into the synchronization, dependency, and reliability of sentiment data between Yahoo and Reddit for each stock. The generated metrics facilitate a deeper understanding of the relationship between these two data sources, contributing to the overall reliability of our findings in the financial analysis domain.

6 DATA VISUALIZATION

6.1 Reddit Politics Posts Activity

In our implementation, we have enriched our data collection strategy by incorporating a dedicated table named "Reddit_Posts_Politics." This table includes essential fields such as post_id, subreddit, and created_utc, providing a structured and organized dataset for a more detailed examination of the r/politics subreddit activity. This meticulous approach ensures that our analysis is not only visually represented through the plotted figure (see Figure 2) but also supported by a comprehensive dataset that captures key attributes of each submission, reinforcing the credibility and depth of our findings.

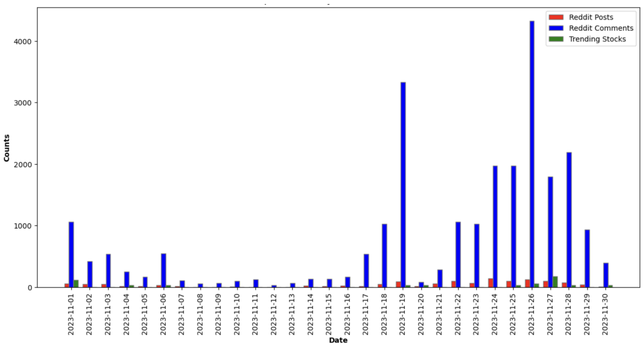


Figure 5: Comparison of Daily Counts for all datasets

6.2 Reddit Politics Comments

In our implementation, we have introduced a dedicated table named `Reddit_Politics_Comments`, featuring essential fields such as `comment_id`, `subreddit`, and `created_utc`. This structured table enhances our ability to conduct a detailed examination of comments within the political discussions on Reddit.

6.2.1 *Data Table: Reddit_Politics_Comments.*

- `comment_id`: Uniquely identifies each comment.
- `subreddit`: Specifies the subreddit, focusing specifically on political discussions.
- `created_utc`: Captures the timestamp of comment creation.

Furthermore, to visually represent the hourly activity of comments, we have plotted a graph (see Figure 3). This graph provides insights into the temporal patterns of engagement within the r/politics subreddit, offering a nuanced understanding of how comments evolve throughout the day. The combination of a well-structured data table and a graphical representation ensures a comprehensive analysis of the Reddit politics comments activity.

6.3 Moderate Hate Speech Activity

In our analysis of hate speech within Reddit discussions, we have implemented a structured approach by creating the `hate_speech` table. This table includes crucial fields such as `post_id`, `comment_id`, `subreddit`, `text_body`, `text_score`, `created_utc`, and `is_hate_speech`. The inclusion of these fields allows us to capture relevant information about posts and comments flagged as potential hate speech, enabling a comprehensive examination of this aspect of online discourse.

6.3.1 *Data Table:* `hate_speech`.

- `post_id`: Uniquely identifies each post associated with potential hate speech.
- `comment_id`: Uniquely identifies each comment associated with potential hate speech.
- `subreddit`: Specifies the subreddit under consideration.
- `text_body`: Contains the textual content of the post or comment.
- `text_score`: Represents the score assigned to the textual content indicating its potential for hate speech.

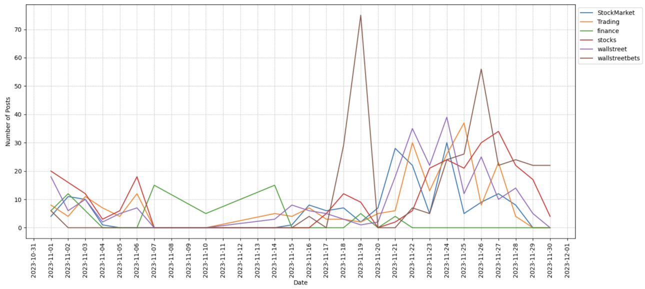


Figure 6: Number of Posts per subreddits

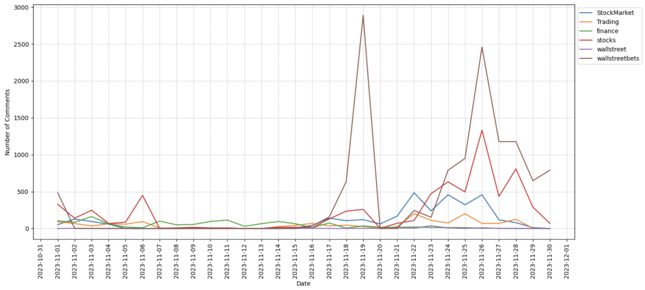


Figure 7: Number of Comments per subreddits

- `created_utc`: Captures the timestamp of post or comment creation.
- `is_hate_speech`: Binary indicator (0 or 1) denoting whether the content is flagged as hate speech.

To complement our analysis, we have plotted a graph depicting the activity related to moderated hate speech (see Figure 4). This graph provides a visual representation of the prevalence and distribution of potential hate speech instances across the specified subreddit. The combination of a well-organized data table and a graphical representation ensures a thorough exploration of hate speech activity within the Reddit community.

6.4 Comparison of Daily Counts for all datasets

- A bar graph (Figure 4) is created to show the amount of data we have got from Yahoo, Reddit Posts and Reddit Comments.

- We are comparing number of Posts, number of Comments, Number of Yahoo stocks altogether by date (Figure 6).

- We are comparing the amount of data by subreddit posts by date and also comparing the amount of data by subreddit comments (Figure 7).

7 CONCLUSION

In conclusion, our project has effectively harnessed Yahoo and Reddit data, offering a comprehensive view of stock sentiments. Analyzing Yahoo data involved meticulous examination of individual stock posts and related news, providing a nuanced understanding of market trends. Sentiment analysis categorized Yahoo posts as negative, positive, or neutral, enhancing our grasp of platform-specific sentiments.

Simultaneously, our exploration of Reddit data considered both posts and comments, tapping into diverse opinions within the community. Sentiment analysis on Reddit comments revealed valuable insights into sentiments circulating on Reddit about various stocks.

The integration of Yahoo and Reddit data in the -`yahoo_reddit_collection` table facilitated in-depth exploration of sentiment synchronization. The subsequent `Ticker_Sync_Data` table, with metrics such as in-sync and out-of-sync instances, dependency percentage, and reliability score, provided crucial indicators of sentiment cohesion and reliability.

Our comparative analysis identified patterns and divergences between Yahoo and Reddit data, enhanced by visualizations. In essence, our project showcased the potential for deriving meaningful financial insights through synergizing diverse data sources. The insights gained pave the way for informed decision-making in the dynamic landscape of financial markets.