Stock Market Social Media Analysis System

Shubhendu Jadhav
State University of New York, Binghamton
Binghamton, New York, USA
sjdhav@binghamton.edu

Avanti Kopulwar
State University of New York, Binghamton
Binghamton, New York, USA
akopulwar@binghamton.edu

Abstract

This proposal outlines the development of a Stock Market Social Media Analysis System designed to collect and analyze data from Reddit and 4chan. The system aims to provide insights into stock mentions, sentiment shifts, meme-driven trading trends, and cross-platform propagation of financial discussions. By leveraging APIs and custom scrapers, the system ensures continuous data collection, enabling comprehensive analysis essential for understanding the dynamics of online financial culture.

CCS Concepts

• **Information systems** → **Data collection and analysis**; • **Social and professional topics** → *Online financial communities.*

Keywords

Stock Market, Reddit API, 4chan Scraper, Sentiment Analysis, Data Science Pipeline

1 Introduction

Online platforms such as Reddit and 4chan play a crucial role in shaping financial discourse. Communities like `/r/wallstreetbets` and 4chan's `/biz/` board have fueled meme stock surges and influenced investor sentiment. Analyzing these discussions provides valuable insights into retail investor behavior, stock popularity trends, and market hype cycles. This proposal presents the design and implementation of a Stock Market Social Media Analysis System, which will collect and analyze data from Reddit and 4chan, laying the foundation for advanced market trend analysis.

2 Description of Data Sources

2.1 Reddit

Reddit is a community-driven platform where stock-related discussions and memes frequently emerge. The Reddit API will be accessed using OAuth 2.0 authentication and custom-built HTTP requests. Targeted subreddits include `/r/stocks`, `/r/wallstreetbets`, `/r/investing`, `/r/pennystocks`, and `/r/options`.

2.2 4chan

4chan's `/biz/` board is an anonymous forum that plays a major role in discussions about cryptocurrencies, meme stocks, and speculative trading. Since it lacks an official API, a custom-built scraper will be used to fetch threads and related posts.

3 Data Collection Plan and System Architecture

3.1 Data Collection Process

Reddit: Posts will be fetched periodically using OAuth 2.0 authentication, collecting titles, text, metadata, and comments. **4chan:** Threads will be scraped every 30–60 minutes from the `/biz/` board, storing posts and associated text.

3.2 System Architecture

The architecture consists of two crawlers (Reddit and 4chan) running in parallel threads, producing JSON snapshots in a structured `data/` directory. Daily aggregated summaries are stored in a `summaries/` folder. Future extensions include integration with MongoDB/Postgres for scalable storage and querying.

3.3 Libraries and Tools

Python `requests`, `json`, `threading`, and `datetime/zoneinfo` will be used for data collection and storage. `Docker` will be employed to containerize the system for portability and easier deployment.

4 Measurements and Analysis Ideas

4.1 Stock Mentions Over Time

Track daily and weekly counts of stock tickers (e.g., GME, AMC, TSLA) across Reddit and 4chan.

4.2 Sentiment Analysis of Stock Discussions

Apply NLP to classify posts as positive, neutral, or negative toward specific stocks.

4.3 Meme Stock Spread Across Platforms

Compare timing and volume of stock mentions on Reddit versus 4chan.

4.4 Market Hype vs. Volume of Posts

Correlate spikes in discussions with real-world stock price or trading volume changes.

4.5 Cultural and Economic Impact

Analyze how online discussions influence investor behavior, particularly around meme stocks and emerging financial trends.

5 Napkin Math

5.1 Reddit

In approximately two days, the crawler collected **27,520 posts** across six cryptocurrency and finance subreddits. This corresponds to about **13,760 posts/day**, projecting to roughly **96,320 posts/week**. The subreddit-level breakdown is:

- `/r/Bitcoin` → 15,890 posts/week
- `/r/CryptoCurrency` → 16,100 posts/week
- `/r/CryptoMarkets` → 15,890 posts/week
- `/r/CryptoMoonShots` → 16,100 posts/week
- `/r/CryptoTechnology` → 16,170 posts/week

- `/r/Ethereum` → 16,170 posts/week

5.2 4chan

From the `/biz/` board, the crawler collected **63,435 threads** in about two days. This averages to **31,717 threads/day**, projecting to nearly **222,019 threads/week**.

5.3 Total Data

The combined volume from Reddit and 4chan is approximately **318,339 entries per week**.

5.4 Data Size

At an estimated **2 KB of JSON metadata per entry**:

- Reddit ≈ 192 MB/week
- 4chan ≈ 434 MB/week
- **Total ≈ 626 MB/week**

Including images and media (averaging **150 KB each** for a fraction of posts), the projected storage requirement grows to **20–30 GB/week**.

6 System Architecture Diagram

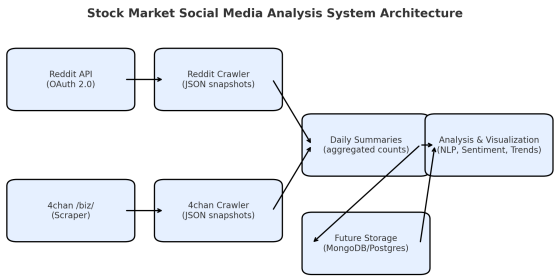

Figure 1: System architecture: the crawlers collect data from Reddit and 4chan, store JSON snapshots, generate daily summaries, and support advanced storage and analysis.

7 Conclusion

The proposed Stock Market Social Media Analysis System provides a robust framework for collecting and analyzing financial discussions from Reddit and 4chan. By implementing continuous data collection and leveraging NLP and visualization tools, the system will yield valuable insights into stock popularity, sentiment shifts, meme-driven trading trends, and cross-platform cultural dynamics.

References

[1] Reddit API Documentation. https://www.reddit.com/dev/api/