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 report describes the Stock Market Social Media Analysis System designed to continuously collect financial discussions from Reddit and 4chan. This system stores historical snapshots, tracks sentiment trends, measures daily discussion volume, and aggregates total logs across platforms. By combining API-based and HTTP-based crawlers with structured storage, the system enables downstream analysis of cross-platform financial narratives, meme-driven hype cycles, and stock sentiment trends.

CCS Concepts

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

Keywords

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

1 Introduction

Online financial communities have become highly influential in shaping stock market sentiment. Platforms such as Reddit's r/finance and 4chan's /biz board generate massive streams of discussions around meme stocks, cryptocurrencies, and retail investor speculation. Understanding these discussions requires continuous data collection, structured storage, and longitudinal analysis. This system collects posts in real time, stores them as JSON snapshots, and aggregates daily trends for future analysis.

2 Data Sources

2.1 Reddit

Reddit offers highly active, topic-specific subreddits that drive retail trader sentiment. We accessed the platform using OAuth 2.0 with custom HTTP requests through the Reddit API. Targeted subreddits included:

- `/r/stocks`
- `/r/wallstreetbets`
- `/r/investing`
- `/r/pennystocks`
- `/r/options`
- `/r/CryptoCurrency`

For each post, the crawler collects titles, scores, timestamps, IDs, subreddit name, and comment metadata.

2.2 4chan

4chan lacks an official API. We used a custom Python scraper to periodically retrieve thread JSON data from the `/biz/` board. Fields collected include thread number, comment body, post ID, thread timestamp, and poster metadata (if available). 4chan is included due to its influence on speculative crypto and meme-stock narratives.

3 System Architecture

The system consists of:

- Two parallel crawlers (Reddit + 4chan)
- Timestamped JSON snapshot storage
- A daily grand-total logger
- Optional Dockerized PostgreSQL storage

Snapshots are stored under `/data/`, while daily summary logs are appended to `/logs/master.log`.

3.1 Architecture Diagram

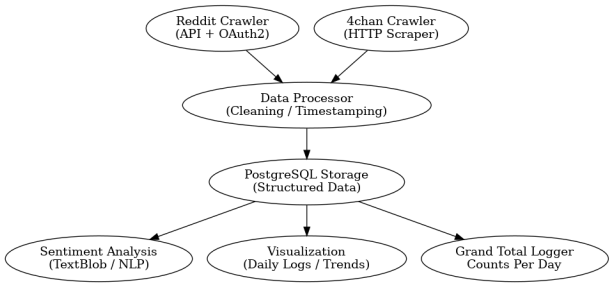

Figure 1: System architecture showing data ingestion, processing, JSON snapshot storage, and future analysis modules.

4 Tools and Libraries

- `requests` — API calls and HTTP scraping
- `json` — snapshot serialization
- `datetime/zoneinfo` — timezone accuracy
- `threading` — parallel crawler execution
- `Docker` — containerized Postgres
- `psycopg2` — database insertion

5 Daily Summary Logging

A dedicated module aggregates:

- total Reddit posts collected
- total 4chan posts collected
- combined total
- timestamp + timezone

This enables longitudinal trend analysis without recomputing.

6 Measurements & Analysis Ideas

6.1 Stock Mention Frequency

Count how often tickers such as GME, AMC, TSLA, NVDA, BTC, ETH appear daily.

6.2 Sentiment Tracking

Future NLP pass will categorize positive, neutral, and negative market sentiment.

6.3 Cross-Platform Propagation

Compare Reddit volume spikes vs. later movement on 4chan.

6.4 Surge Detection

Identify unusual hype patterns before earnings or headlines.

7 Visualization of Data Growth

To better understand how discussion volume evolved during the collection period, we plot the cumulative number of posts retrieved from Reddit and 4chan over time.

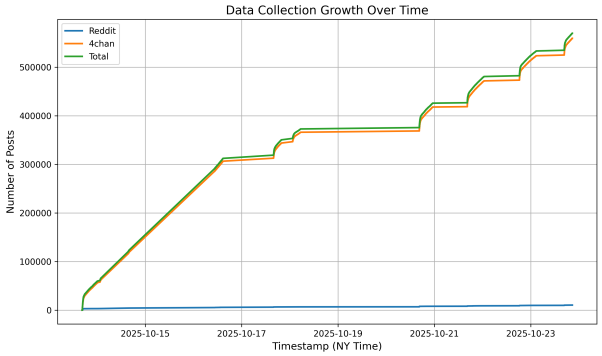

Figure 2: Cumulative growth of Reddit and 4chan posts over time. The total curve closely follows 4chan due to significantly higher posting volume on `/biz/`.

Figure 2 shows that 4chan contributes the majority of collected posts, resulting in steep step-like jumps corresponding to periods of heightened discussion around cryptocurrency and meme stocks. Reddit grows more gradually due to API rate limits and lower daily volume.

8 Napkin Math

8.1 Reddit

In 2.5 days of continuous crawling, 3,316 posts were collected. Projection: ~930 posts/day ⇒ ~6,510 posts/week.

8.2 4chan (/biz/)

In the same period, 62,432 posts were collected. Projection: ~24,900 posts/day ⇒ ~174,300 posts/week.

8.3 Combined Weekly Volume

Total ≈ **180,810 posts/week**.

8.4 Storage Size

At 2 KB metadata per entry: ≈ 360 MB/week JSON data. With media expansion: 5–8 GB/week.

9 Limitations

- 4chan anonymity reduces user tracking fidelity
- Reddit API rate limits require batching
- Sentiment sarcasm can skew NLP

10 Conclusion

This Stock Market Social Media Analysis System provides reliable, timestamped data collection from two influential financial communities. Its modular structure supports future extensions such as database indexing, NLP pipelines, and predictive analytics. The daily summaries and snapshot storage enable long-term research into investor psychology, meme-driven hype cycles, and cross-platform narrative flow.

References

[1] Reddit API Documentation. https://www.reddit.com/dev/api/
[2] 4chan API Unofficial Spec. https://github.com/4chan/4chan-API
[3] RFC 6749: OAuth 2.0 Authorization Framework. https://datatracker.ietf.org/doc/html/rfc6749