GEORGIA INSTITUTE OF TECHNOLOGY, DATA AND VISUAL ANALYTICS PROJECT

Mining Reddit for Stock Valuation Signals

Matthew Mackowski
mmackowski3

Oscar Goodloe
ogoodloe3

Daniel Villaveces
dvillaveces3

Juan Antonio Garcia
jgarcia353

Xu Anne Zhang
xzhang3008

Jake Robertson
jrobertson74

Introduction

Vast amounts of information easily available at the fingertips of any internet user has left enduring changes on the real world - and the stock market is not immune. In early 2021,a social media storm caused the Gamestop stock to go viral, resulting in dramatic swings in stock price. It is events like this that have inspired us to investigate how to derive market insights from social sentiment. This information would be advantageous to anyone involved in stocks. As such, we have created an interactive dashboard for users to easily explore stock price behavior in relation to current events and trends.

Data Gathering and Analysis

- We obtained data using the Reddit API, and overcame the limits on parent posts by collecting child comments recursively, allowing us to retrieve many comments in a short range of time
- We determined how to apply sentiment analysis within the social constructs of Reddit, where users can also vote in favor of or against each post
- The data collected was 400-500MB of individual reddit text, upvotes, timestamp, and author data
- We elucidated major topics of discussion of the Reddit comments by timeframe along with the stocks mentioned
- We deciphered when a comment can be interpreted as a positive or negative stock signal
- We iteratively evaluated and revisited methods used to evaluate how well they achieve our objectives and function in scalability, accuracy, and usability

INTERACTIVITY

Users can change both the selected ticker and the date range to dynamically update visualizations and modeling. LDA has a fixed date range for efficiency reasons and practicality – most useful signal from global stock discussions is gone after a few days.

A HOLISTIC VIEW

Sentiment for the selected ticker can be viewed directly alongside returns.

REDDIT IDIOSYNCRASIES

We account for some of the main features of Reddit in our visualizations. Different communities have different perspectives and areas of focus. A user can visualize sentiment by different subreddits, and our formula for weighted sentiment has signal from Reddit's voting system baked in.

Results

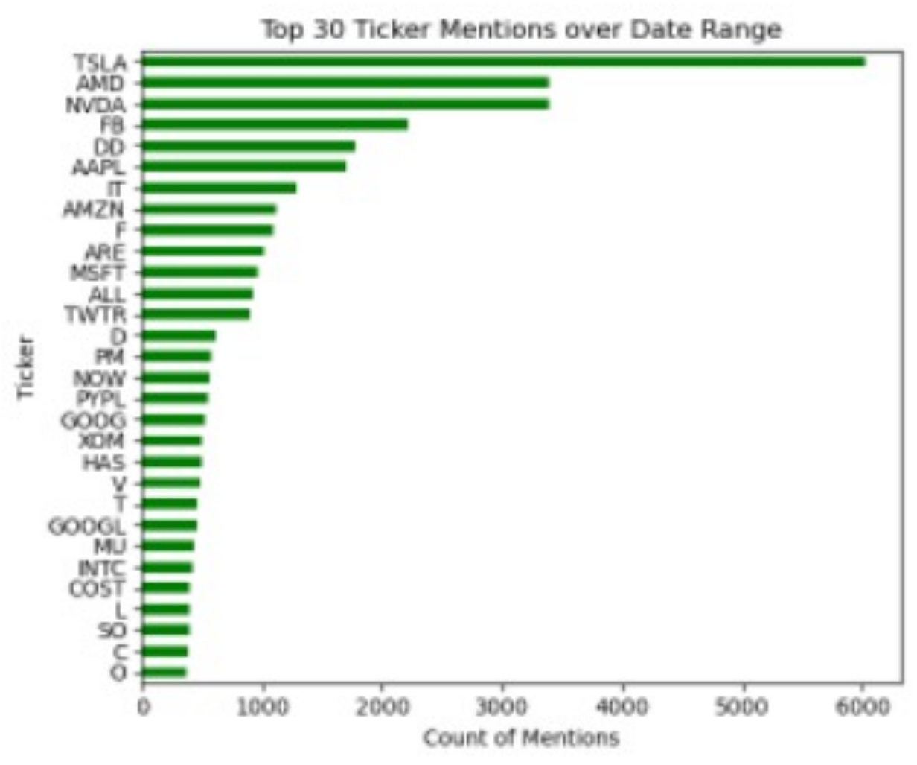
LDA: Relevant topics for last 3 days

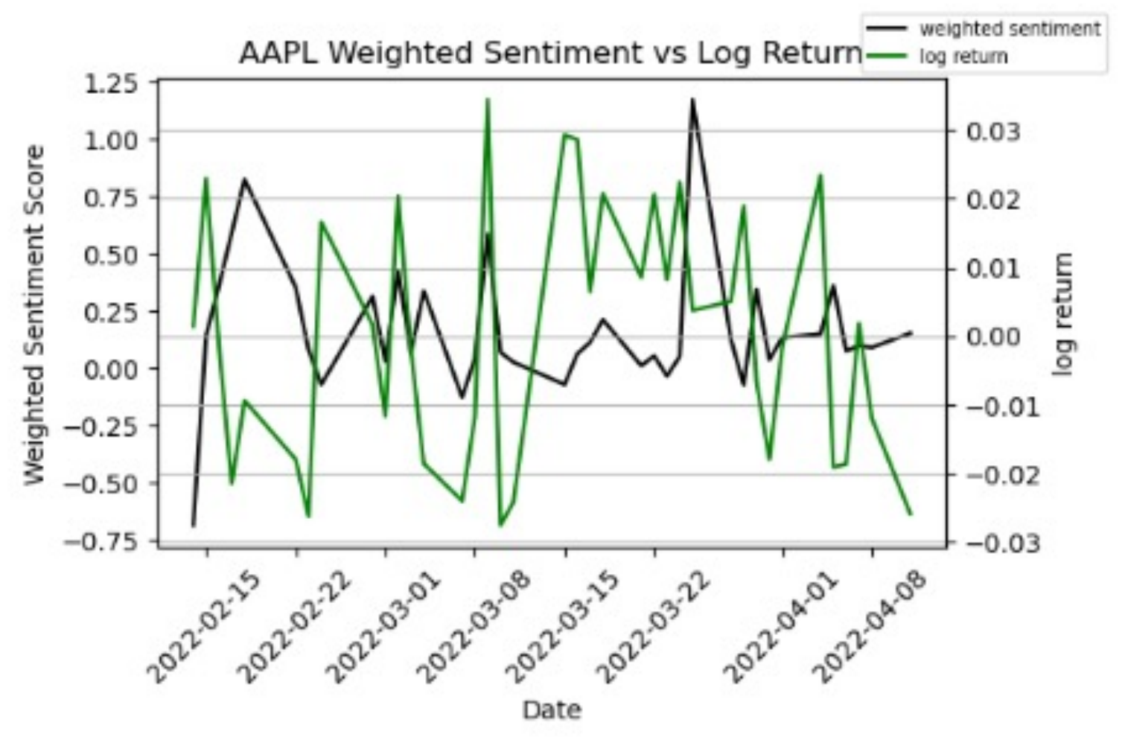
	0	1	2	3	4	5	6	7	8	9
Topic 1	crypto	bitcoin	btc	see	still	even	much	also	way	use
Topic 2	puts	calls	spy	shit	day	back	year	cpi	today	fuck

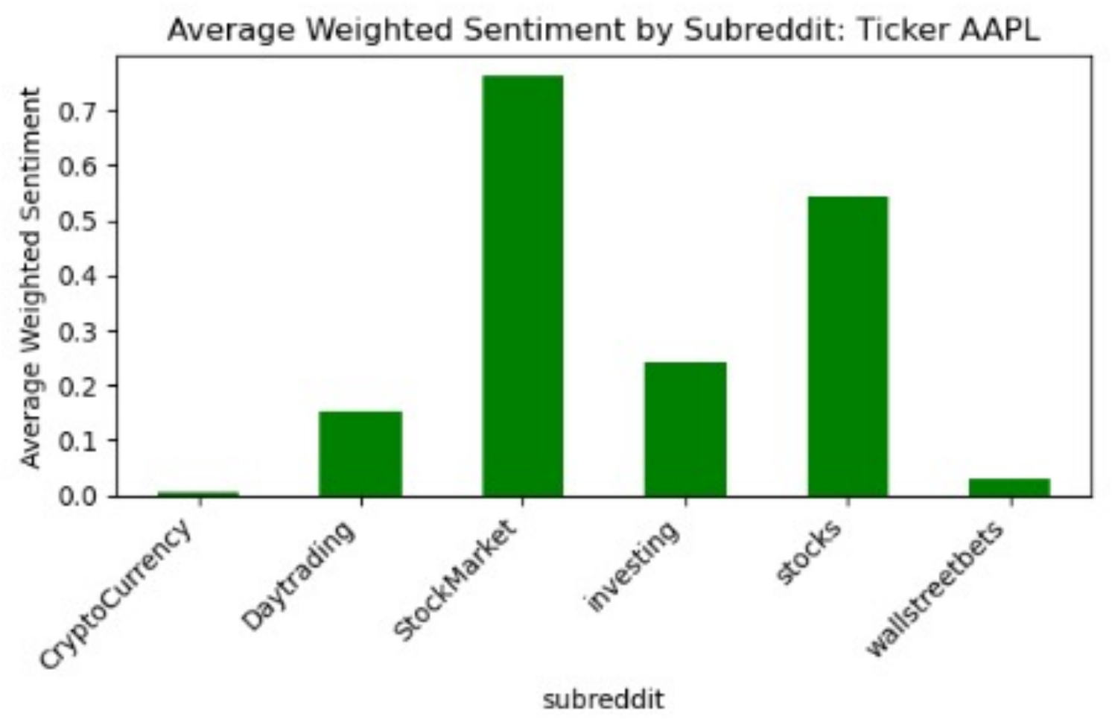
Stock Ticker: AAPL

date_range 02/13/22-04/12/22

Symbol	Name	Sector
AAPL	Apple	Information Technology







```
Binary Next Day Increase Logistic Regression Model
Next Day Prediction:  [1]
```

	0	1	accuracy	macro avg	weighted avg
precision	0.528736	0.512442	0.513652	0.520589	0.520436
recall	0.080000	0.931323	0.513652	0.505662	0.513652
f1-score	0.138973	0.661118	0.513652	0.400045	0.404946
support	575.000000	597.000000	0.513652	1172.000000	1172.000000

INTERACTIVE DASHBOARD

SEE GLOBAL TRENDS

LDA informs the user of the most pressing points of current discussion. The user can inform themselves without reading hundreds of posts and comments.

EASE OF ACCESS

For the selected ticker, the full company name and sector are shown. Even users relatively unfamiliar with the securities market can use this tool to aid decision making.

PREDICTIVE MODELING

The dashboard automatically updates the model for the selected ticker. Simple prediction included for accessibility, with model characteristics displayed for advanced users.
Puts powerful predictive capabilities in the hands of non-experts.

Methodology and Experiments

We used TextBlob models, post data, and a variety of classification models in our analysis to intuitively identify patterns. The dashboard we have created provides a series of clear graphs allowing users to quickly toggle different filter settings to find insights in subsections of data. The power of this is the combination of various components in a single place for ease of use and exploration.

The main methods and how they work include:

- Sentiment Analysis
  - Using Python TextBlob, we produced a polarity rating of sentiment and subjectivity for any given comment. We decided to also incorporate community score data directly from Reddit to determine the level of support each comment receives.
- Topic Modeling
  - We used Latent Dirichlet Allocation and text pre-processing. We addressed repeated words, how to interpret the raw form, and included n-grams (n-length sets of words passed as single entries) to consider phrases coherently.
- Logistic Regression
  - We modelled the stock price activity as it corresponded to Reddit post data and sentiment analysis and compared this to real stock prices at later points in time.

Conclusion and Summary

Our project holistically blended topic modeling, social media sentiment analysis, and predictive modeling of security performance to identify the impact that world events and investor sentiment have on investment markets. Beating the market is the dream, and the information derived in our multi-visual dashboard may be used as a quantifiable way to achieve this.

What was new? The innovations in our approach include:

- Real-time calculations and filtering of data
- Assembling of multiple pieces from logistic regression, sentiment analysis, topics derived from LDA, together in a coherent visualization
- Gives social context to investors and those interested in the stock market
- Creative use of machine learning techniques to automatically extract a large amount of comment topics