Data Science Capstone - Final Presentation

Predicting Stock Market Movements through Social Media Sentiment Analysis

Authors: Aryan Bhardwaj, Tyler Gomez Riddick

August 13, 2024

TABLE OF CONTENTS

Introduction

Introduction

- The Dynamic Finance World:
  - Traditional stock analysis methods are now being supplemented by data-driven approaches.
- The Power of Social Media:
  - Social media platforms, especially X and Reddit, have emerged as significant sources of market sentiment, where opinions and reactions are expressed in real time.
- Significance of Sentiment Analysis:
  - By analyzing sentiment on platforms like X and Reddit, we can gain insights into trends, offering a potential predictive edge in stock movements.
- Project Goal:
  - Analyze social media posts and correlate them with financial data from Yahoo Finance to develop a predictive model for stock market movements.

Background

Background

Rationale for Social Media Sentiment Analysis:

- Platforms influence public opinion and investor behavior.
- Traditional financial analysis often misses the immediate trends discussed on social media

Why Focus on These Companies?

- Nvidia & Wingstop: Nvidia's and Wingstop's growth have made both popular subjects on social media, ideal for sentiment analysis.
- GameStop: The Reddit-driven short squeeze exemplifies social media's power in influencing stock prices.
- Dogecoin: Propelled by social media and endorsements from figures like Elon Musk

Methods

Three Phases

Web Scraping

Sentiment Analysis

Modeling

Overview

Data Collection:

- Gather stock price data from Yahoo Finance and social media data from platforms like X and Reddit through web scraping.

Sentiment Analysis:

- Analyze social media posts to quantify sentiment and gauge public opinion on the selected companies.

Model Development:

- Build predictive models using the sentiment data combined with historical stock prices to forecast market movements.

Web Scraping

- Twitter
  - Goal: scrape top 250 posts for each day in date range
    - API expensive
    - Strict rate limits
  - twscrape
  - Custom Selenium-based web scraper

- Reddit
  - Goal: scrape all submissions in date range
  - Python Reddit API Wrapper (PRAW)

Web Scraping

- twscrape
  - Python package
  - Uses official Twitter API
  - Easy to deploy
  - Collects data very quickly
  - Easy to hit rate limits
  - Accounts banned quickly

Twitter/X

- Selenium
  - Custom built using Selenium package
  - Cycles through short list of accounts
  - Single-threaded
  - Reliable, and can be trusted to run for hours without fail
  - Significantly slower than twscrape

Web Scraping - PRAW

- Reddit API
  - Free to sign-up for and use
  - Access to vast amounts of data
  - High rate limits
  - Accessed via PRAW
- Six subreddits
  - r/gamestop
  - r/wingstop
  - r/nvidia
  - r/dogecoin
  - r/GME
  - r/NVDA_Stock

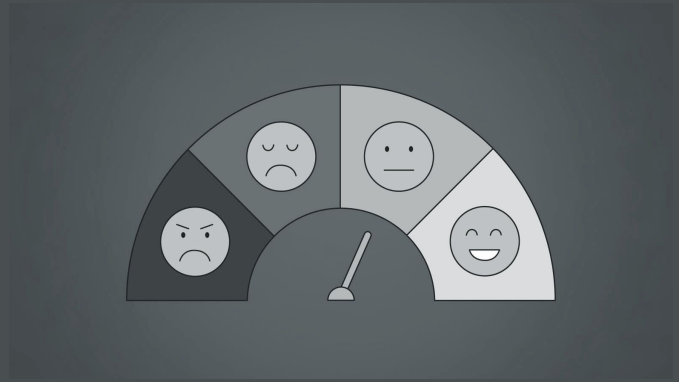
Web Scraping - Ethical Considerations

- Rate Limits
  - Developers impose rate limits to avoid undue burden
  - Data is publicly accessible
  - Never exceeded a rate limit using Selenium
- Data Anonymization
  - Data and users linked
  - No user data was collected at any point

Sentiment Analysis

Sentiment Scoring Process

- Calculated sentiment score for each post.
  - vaderSentiment Python package
  - Tuned to social media sentiments
- Aggregated sentiment scores daily via averaging to align with stock market data frequency.

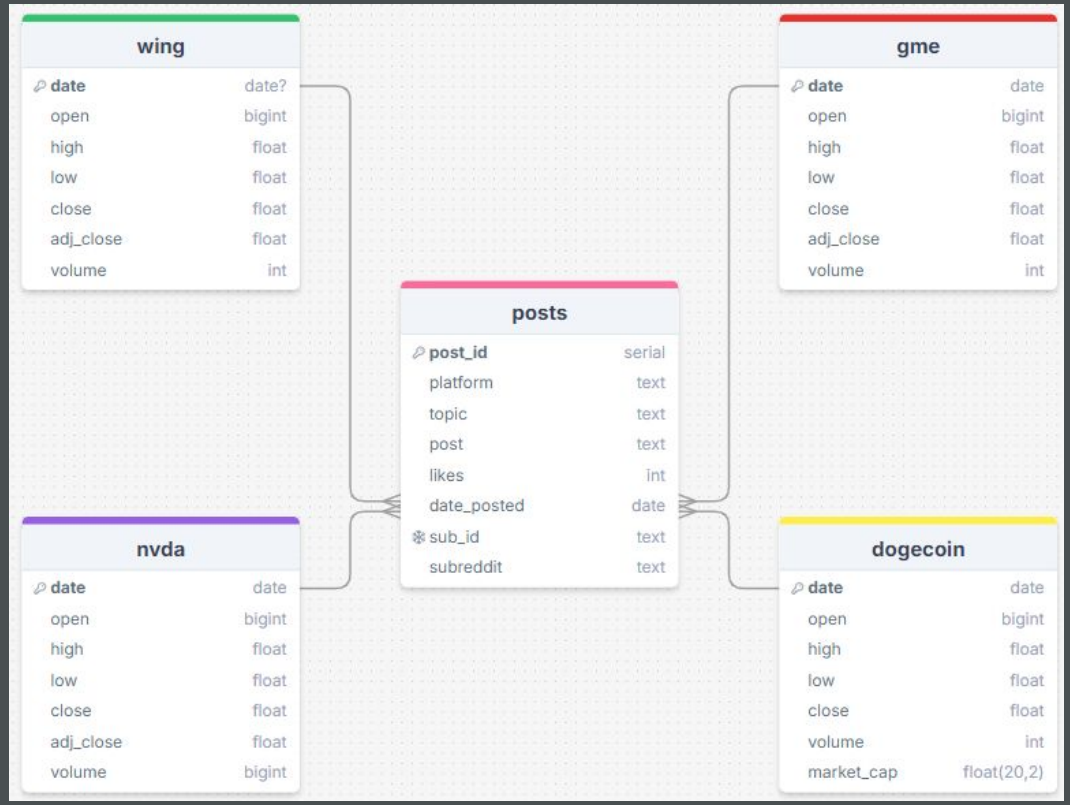


Modeling

- Goal: to predict whether the closing price for a stock entity will move up or down during the following days depending on the social media sentiment about that entity
- (Primarily) a binary prediction: “up” if the closing price is higher the next day or “down” if it is lower
- Predicting next day change, as well as next four days after that
- Two models
  - Logistic regression
  - Decision trees
- Using scikit-learn package
- Measuring performance by model accuracy

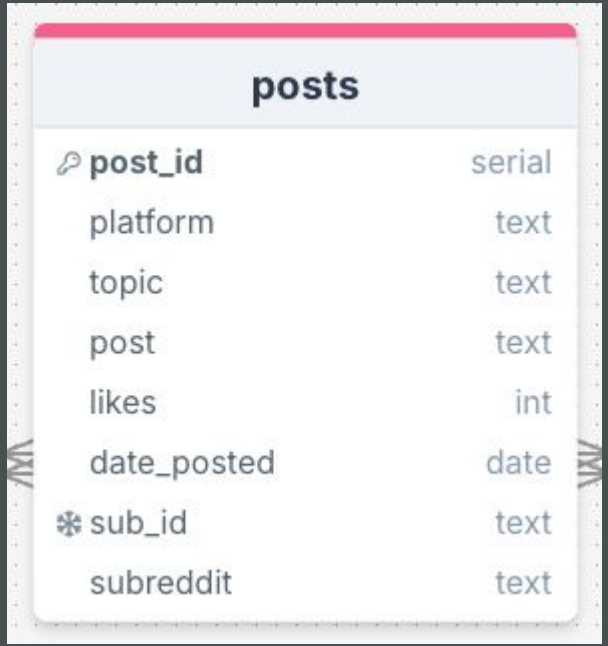
Data

Database


wing
date date?
open bigint
high float
low float
close float
adj_close float
volume int
gme
date date
open bigint
high float
low float
close float
adj_close float
volume int
posts
post_id serial
platform text
topic text
post text
likes int
date_posted date
sub_id text
subreddit text
nvda
date date
open bigint
high float
low float
close float
adj_close float
volume bigint
dogecoin
date date
open bigint
high float
low float
close float
volume int
market_cap float(20,2)

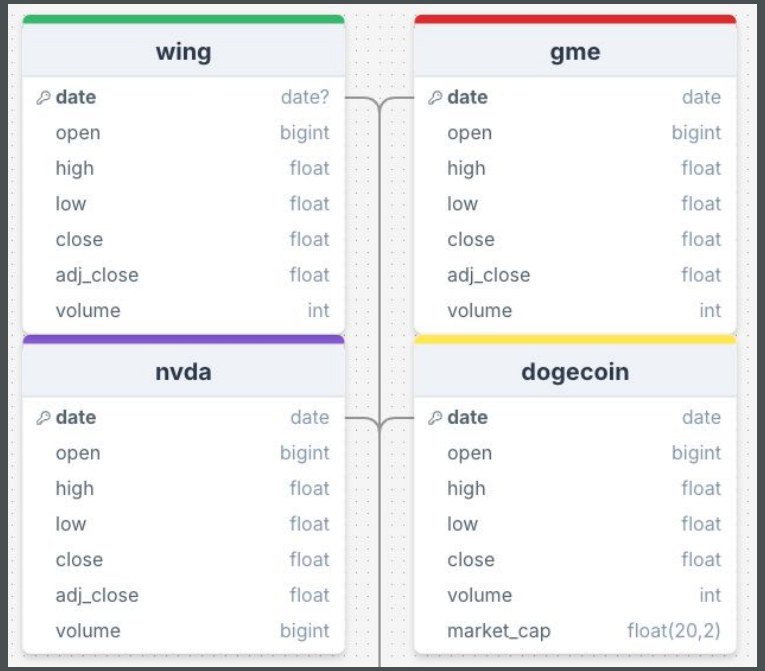
Social Media Data - Posts

- Relevant columns:
  - **post** - the actual text content of each X post or submission
  - **likes** - number of likes (X) or net score (Reddit)
  - **date_posted** - the date the post was added to the site
    - Foreign key to the stock market table



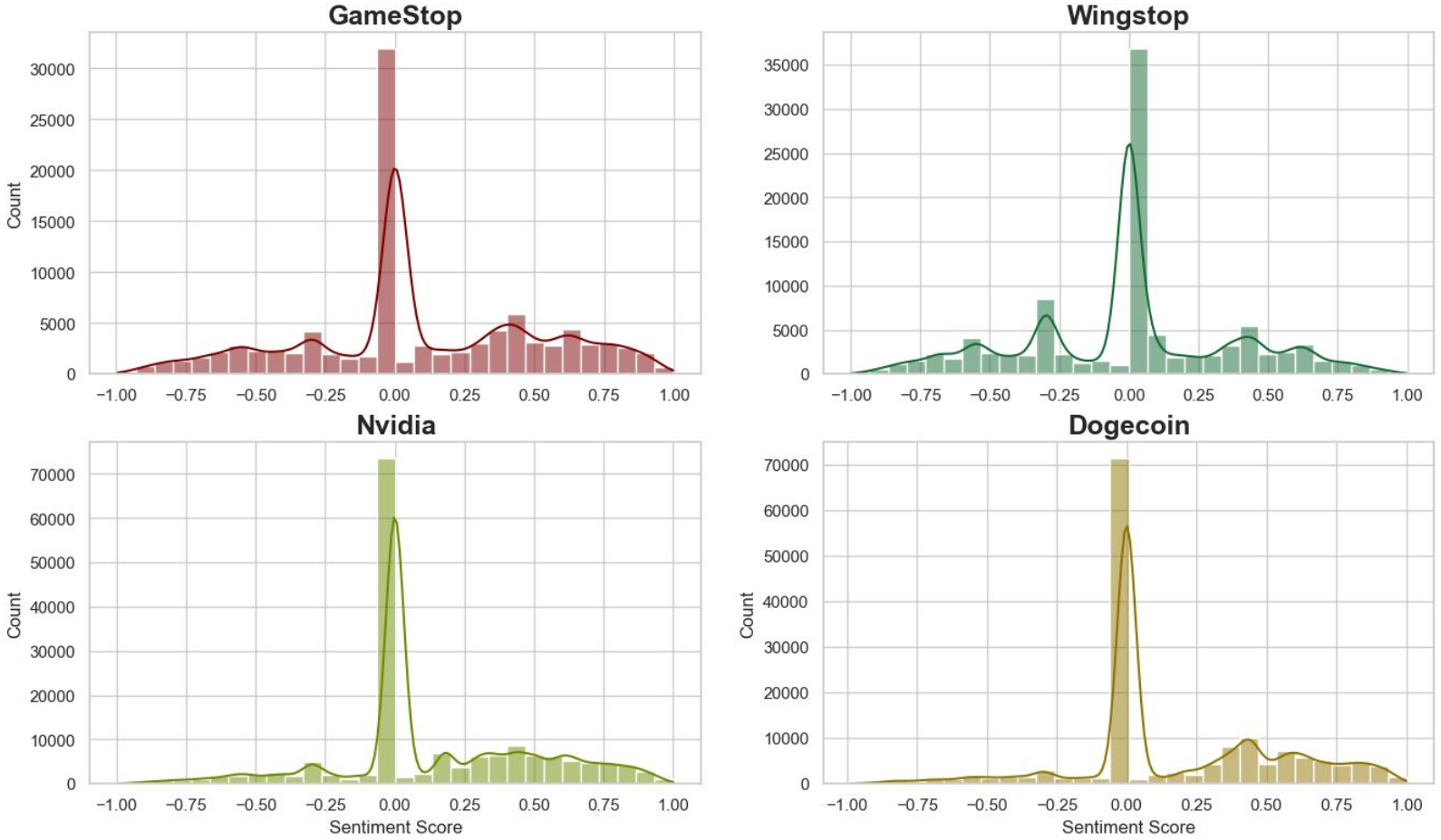
Stock Market Data

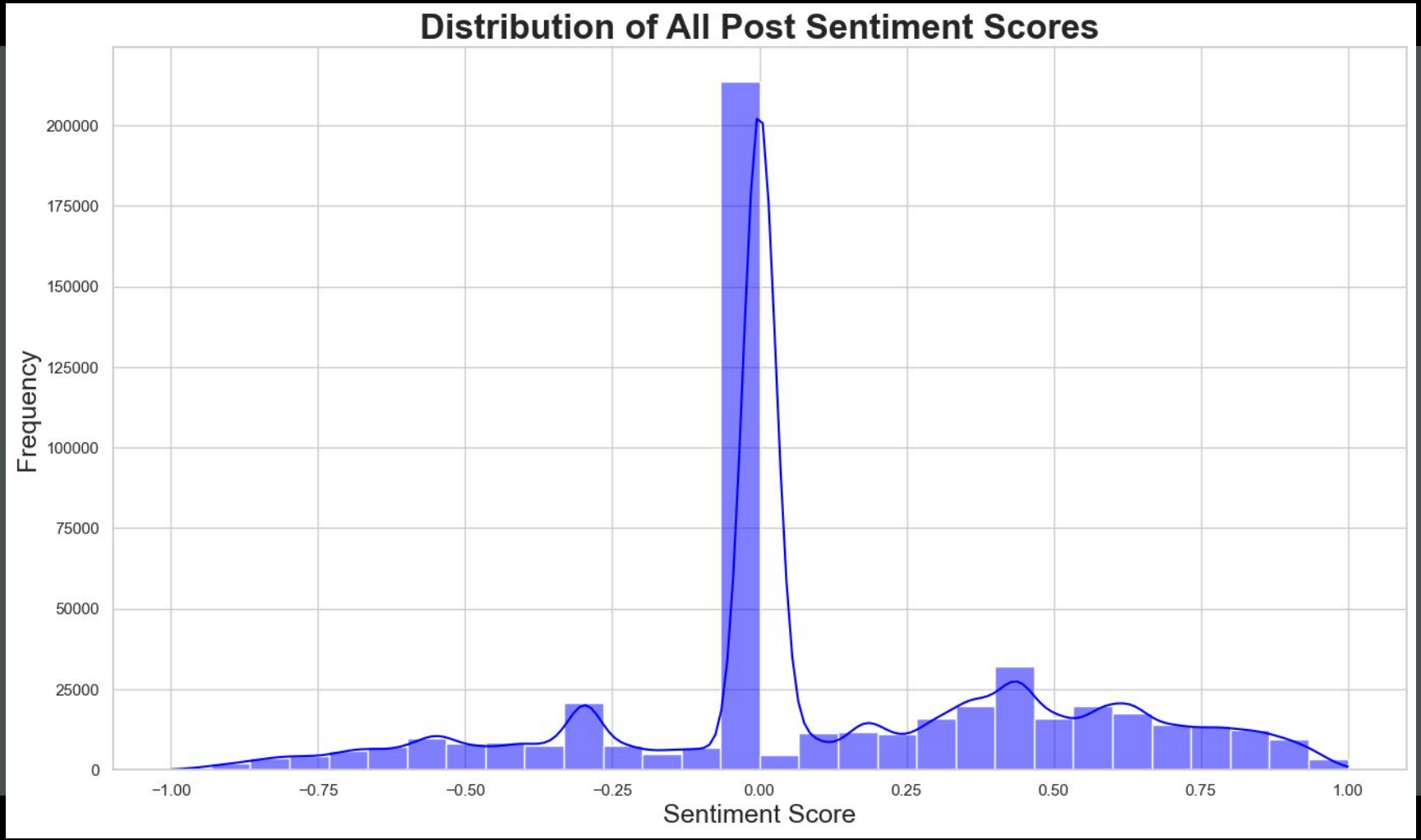
- Most relevant columns
  - **date** - primary key
    - the market date
  - **open** - the stock price at the start of the market day
  - **close - the stock price at the end of the market day**
  - **volume** - the number of stocks traded (bought or sold) during the market day
  - For wing, gme, and nvda:
    - **adj_close** - closing price adjusted for actions taken by company that would affect price
  - For dogecoin:
    - **market_cap** - market capitalization, total value of all outstanding shares



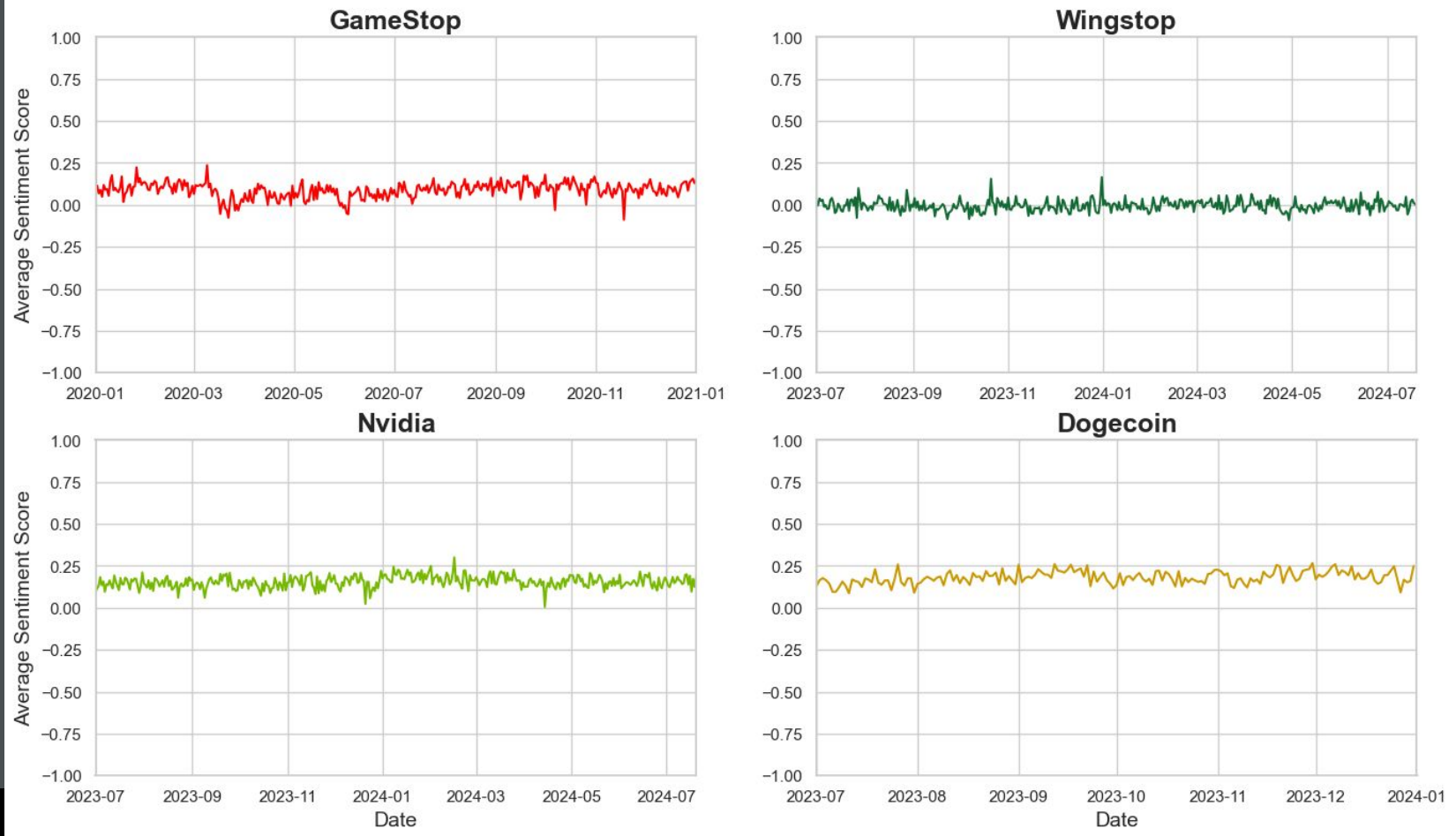
Results

Distribution of Post Sentiment Scores

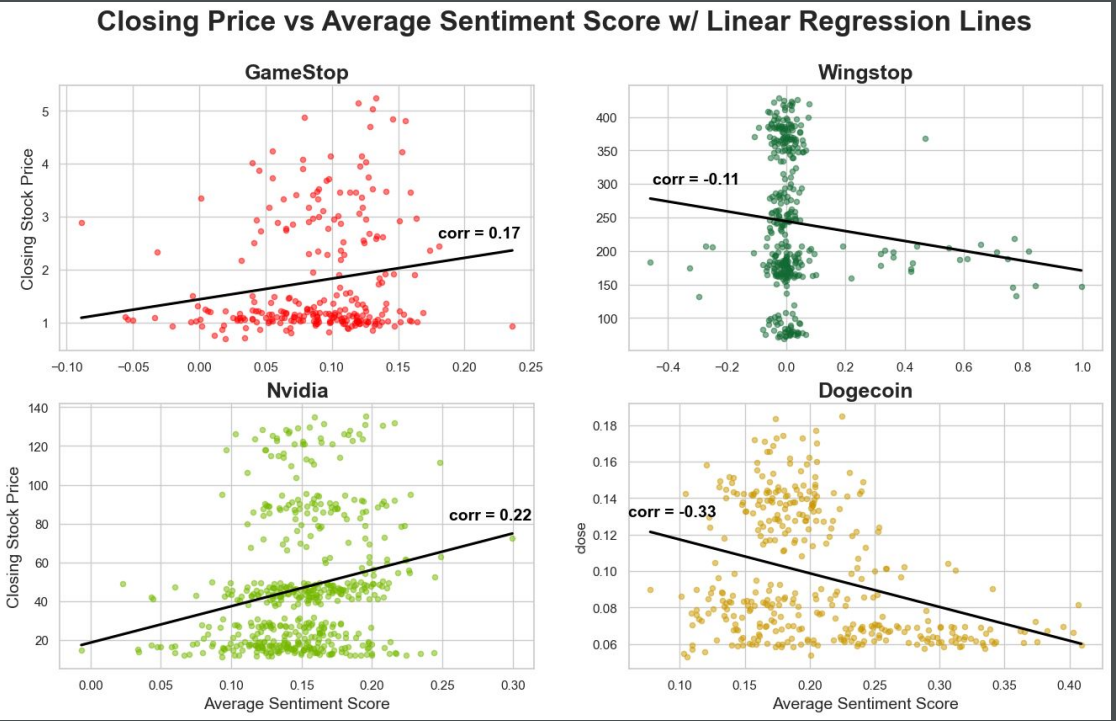



Distribution of All Post Sentiment Scores
Frequency
200000
175000
150000
125000
100000
75000
50000
25000
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment Score

Average Daily Sentiment Score vs Time

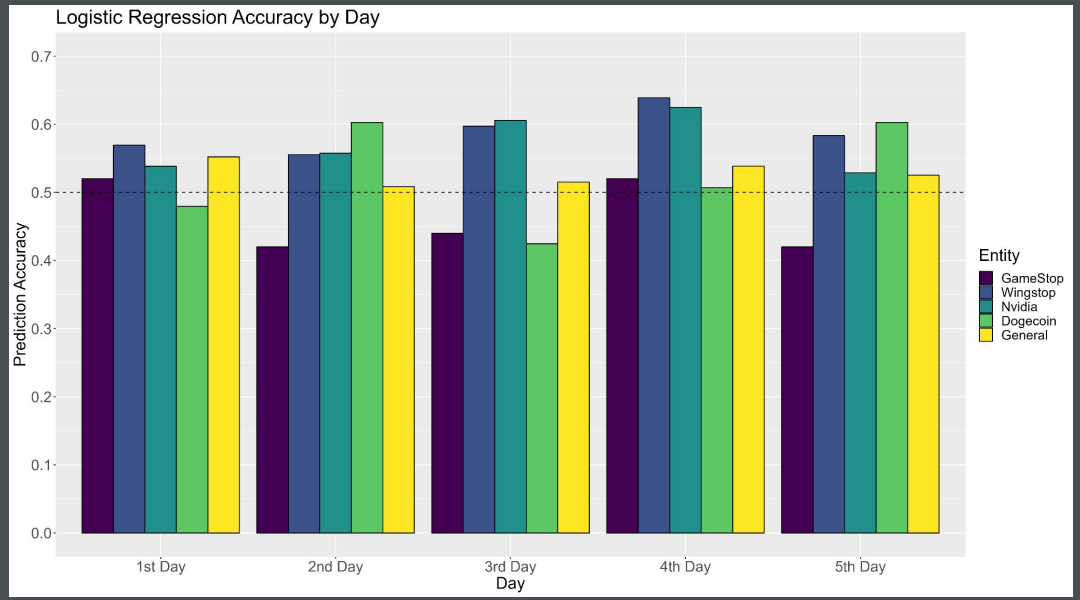


Correlation



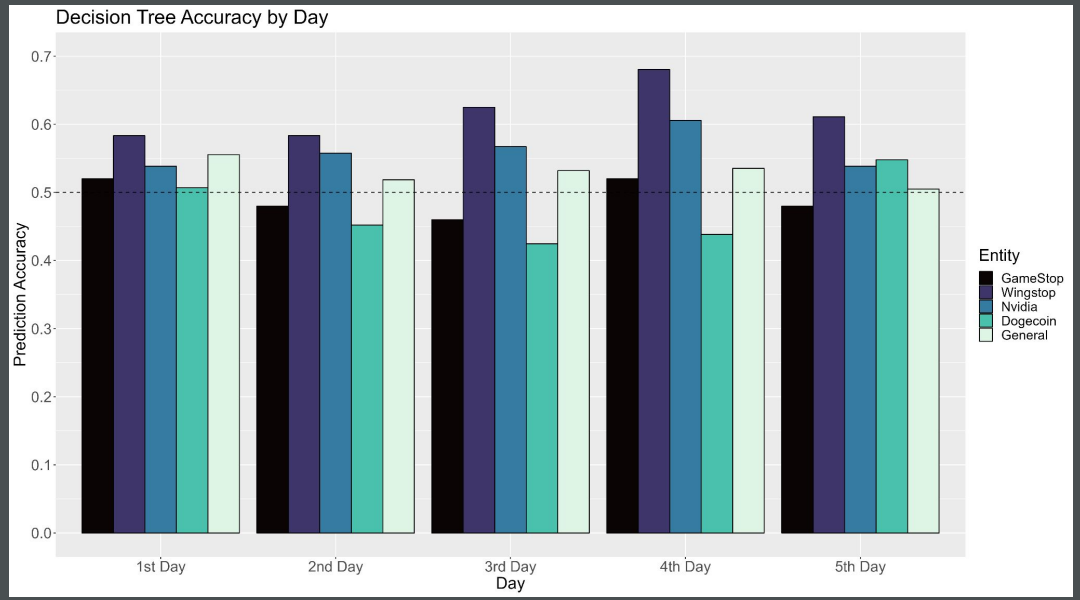
Logistic Regression Model

- **Wingstop** and **Nvidia** models perform best
- **Gamestop** and **Dogecoin** display varying accuracies
- **General** model is consistently above 50%



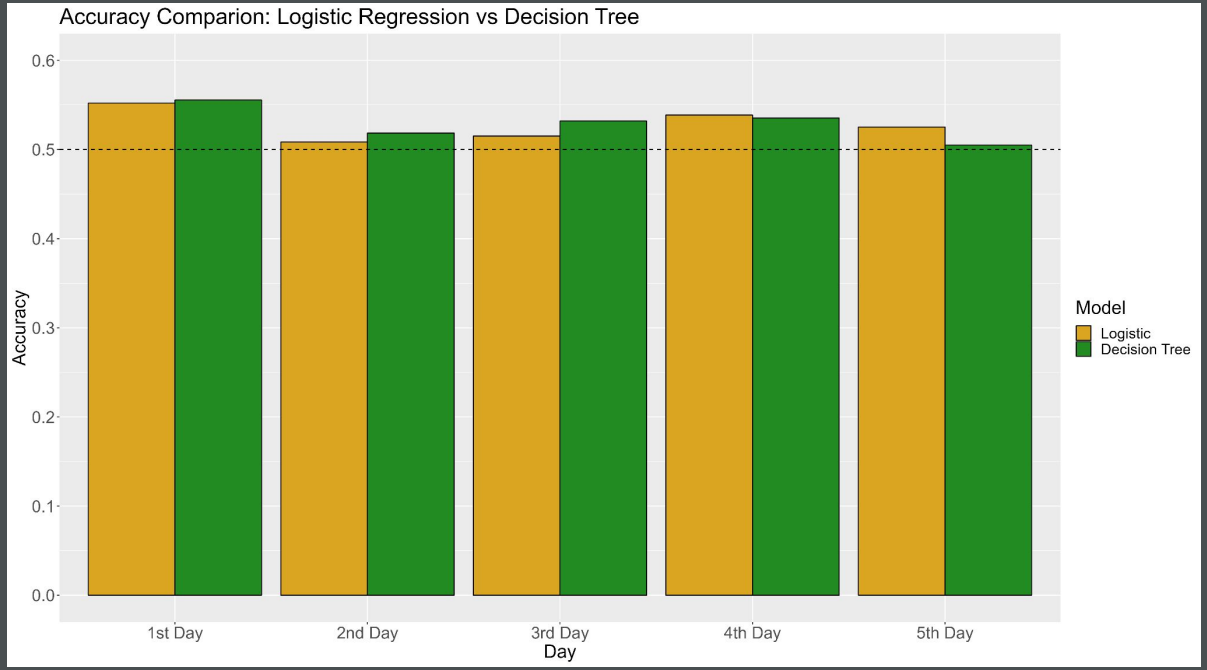
Decision Tree Model

- **Wingstop and Nvidia** models perform best
- **Dogecoin** accuracy improves
- **GameStop** improves on days where logistic model predicted with < 50% accuracy
- **General** model is consistently above 50%



Comparison

- Decision tree performs better on first three days
- Logistic regression performs better on the 4th and 5th days
- These results are very close across the board though



Comparison - McNemar Test

- Results are statistically significant for all days except the second day
- Decision Tree model is the more accurate model on the 1st and 3rd days
- Logistic Regression model is more accurate on the 4th and 5th days

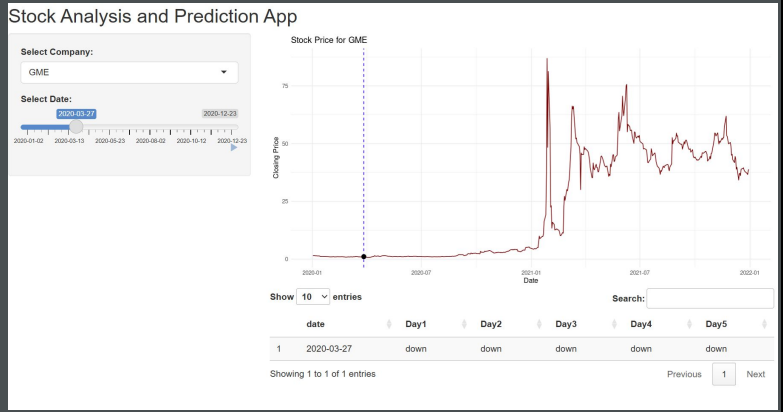
Day	P-Value
1st Day	$\mathbf{9.54 \times 10^{-7}}$
2nd Day	**1.0**
3rd Day	$\mathbf{4.34 \times 10^{-19}}$
4th Day	$\mathbf{2.27 \times 10^{-13}}$
5th Day	$\mathbf{1.22 \times 10^{-3}}$

Conclusion

Insights

- Average daily sentiment does appear to be moderately predictive of daily change for these four entities
- Entities with higher correlations (positive or negative) tended to have higher accuracies in both models
- Decision tree seems to be more predictive for immediate days
- Logistic regression is more predictive 4 to 5 days out from date of interest
- Stock Analysis and Prediction App -

  Stock Analysis and Prediction App (shinyapps.io)



Future Work

- Incorporating additional features
- Better time resolution (hourly vs daily)
- Wider dataset; incorporating more stock entities
- Incorporating broader market factors for a better understanding
- Enhancing dashboard with live, hourly data

Questions?