Sentiment Analysis on r/WallStreetBets and predicting the success of their market strategies

CSE342 - Statistical Machine Learning Course Project

By Saksham Singh (2022434), Ritika Thakur (2022408)

Problem Statement

In recent years, social media platforms such as Reddit's r/WallStreetBets have emerged as influential forums for discussions and opinions on stock investments. However, analyzing the vast amount of unstructured data generated by these communities, not to mention their colorful and profane jargon, and aggressive trading strategies, it presents a significant challenge for investors seeking to make informed decisions. In this project we created a Stock Prediction Model and did sentiment analysis on posts of the Reddit r/WallStreetBets.

Problem statement - Sentiment Analysis

Develop a sentiment analysis model capable of accurately assessing the sentiment expressed in posts and comments on r/WallStreetBets. The model should be trained to differentiate between positive, negative, and neutral sentiments regarding specific stocks mentioned in the forum.

Problem Statement - Stock Prediction

Design and implement a machine learning model for predicting the future value of stocks based on historical price data, market indicators, and sentiment analysis results. The model should be capable of generating accurate predictions over various time horizons.

Dataset used

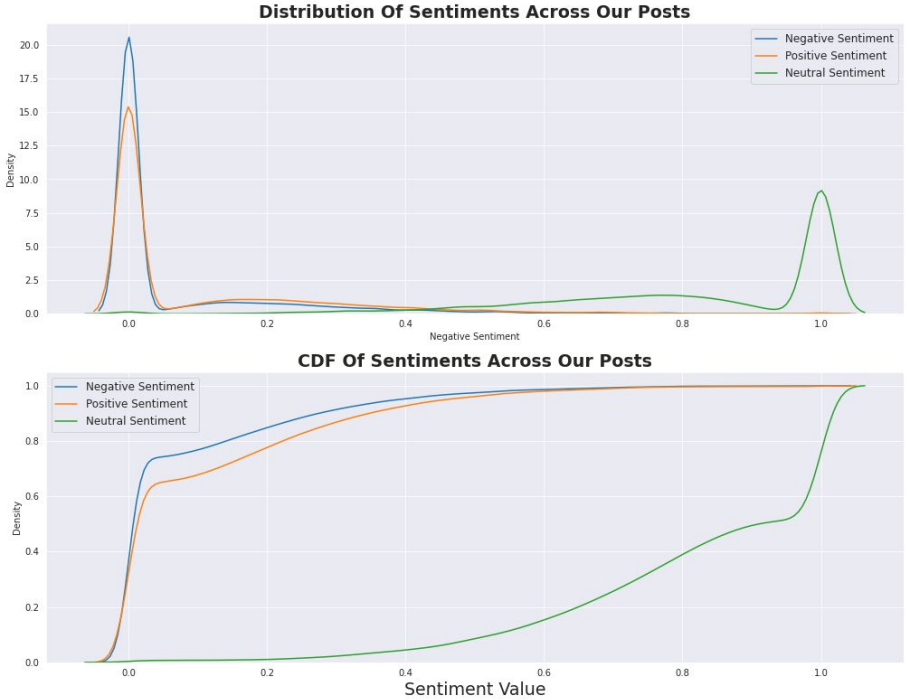
- **Reddit-WallStreetBets-Posts -** a dataset on Kaggle featuring 53K unfiltered posts from the subreddit r/WallStreetBets dating from 29-09-2020 to 16-08-2021. The dataset contains post title, post body, posting timestamp, and the score (upvote-downvote sum) of the post.
- **Yahoo Finance -** Using YFinance python library we were able to get the data of nearly every major listing on Nasdaq, with the open price, the high, the low, and the closing price on every trading day since recorded history.

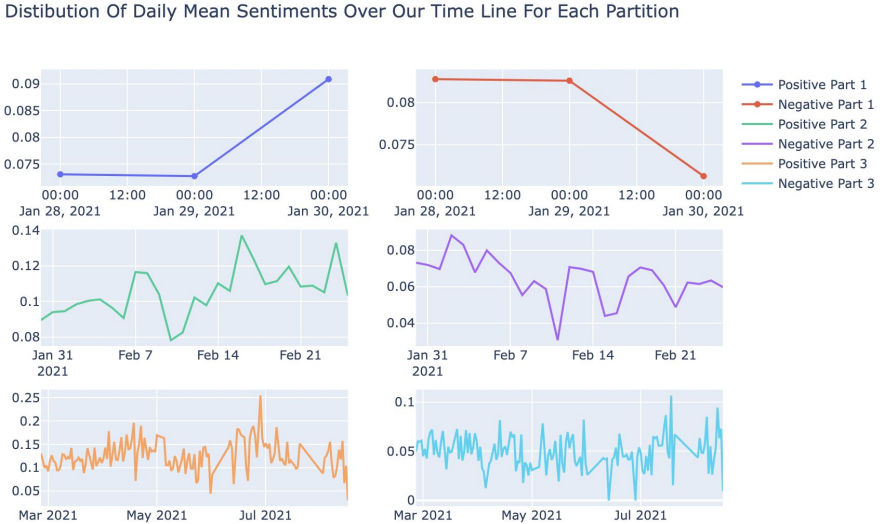
Proposed Architecture - Sentiment Analysis

VADER(Valence Aware Dictionary and sEntiment Reasoner) is good at understanding sentiment in social media text, like tweets and online comments. It works by looking at a list of words with positive or negative connotations, and then considering things like capitalization and punctuation to get a more nuanced understanding of the feeling behind the text. We utilised VADER to give a sentiment score to each post's title and body and determine the sentiment as positive, negative or neutral.

We cleaned our text (text preprocessing), utilized VADER to determine the sentiment score for each text and extracted relevant features(feature engineering) to visualise the result.

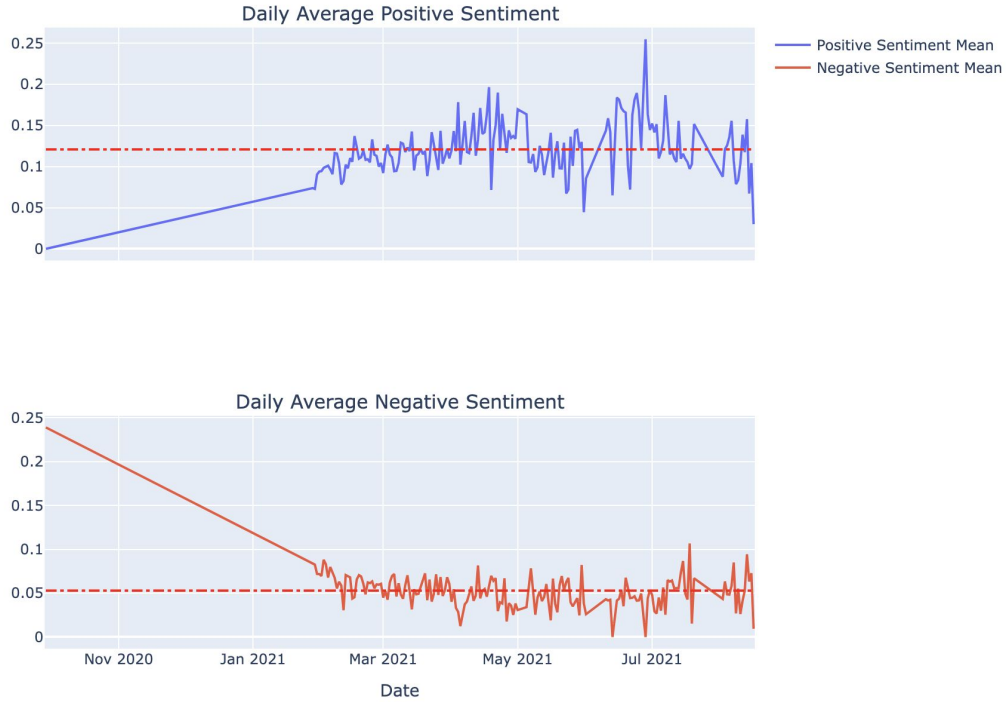
Analysis on Post Titles

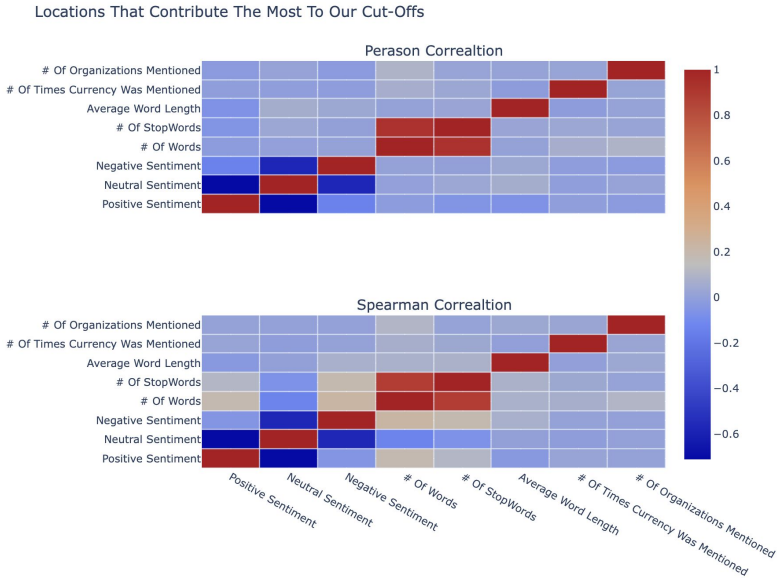




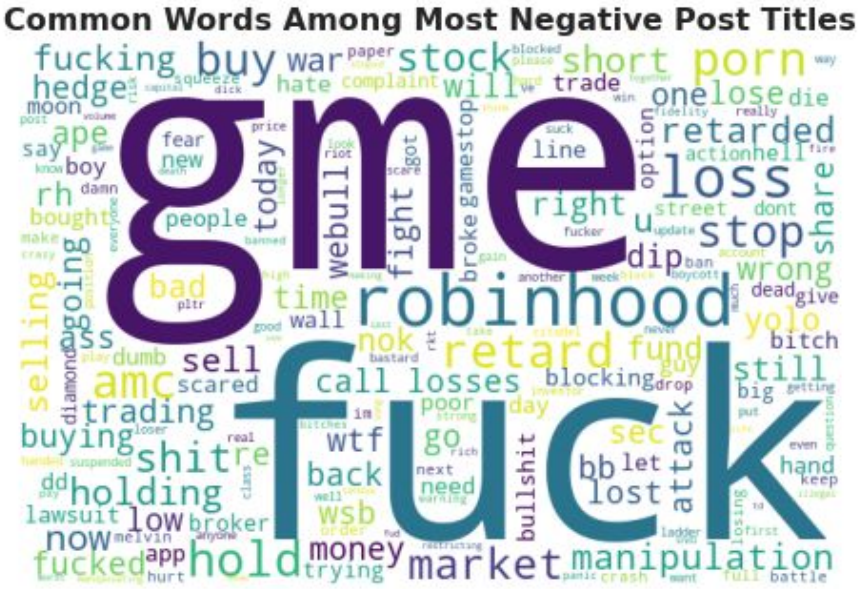
Analysis on Post Titles

Sentiment Average Change With Time



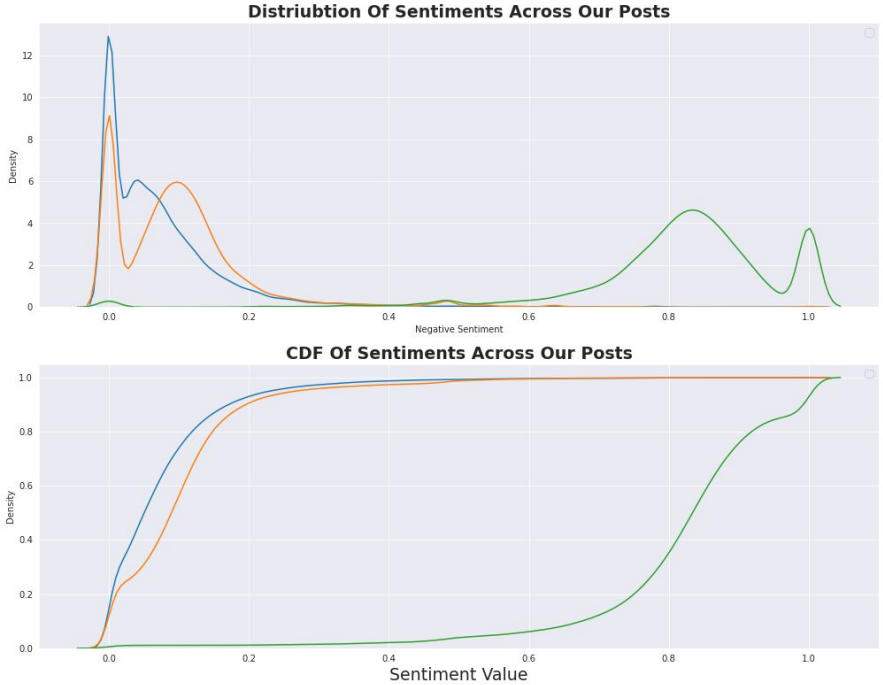


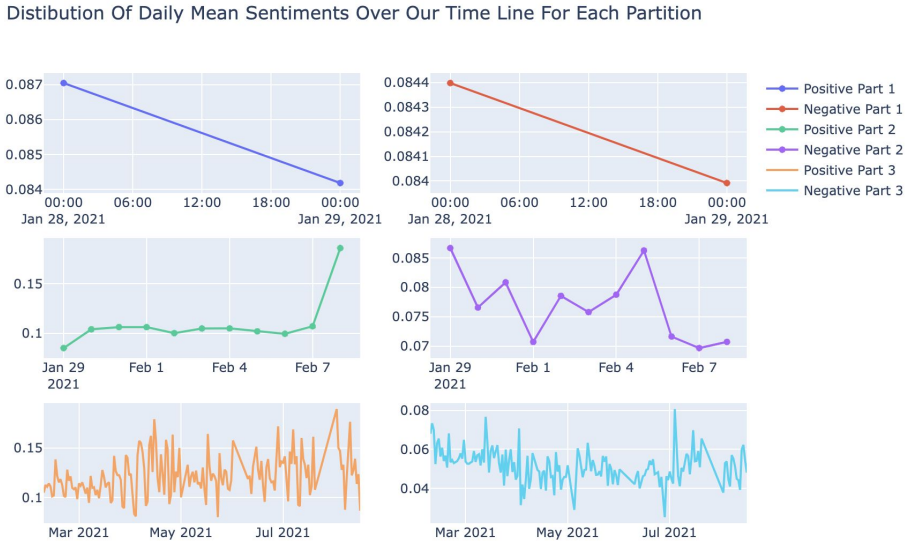
Analysis on Post Titles



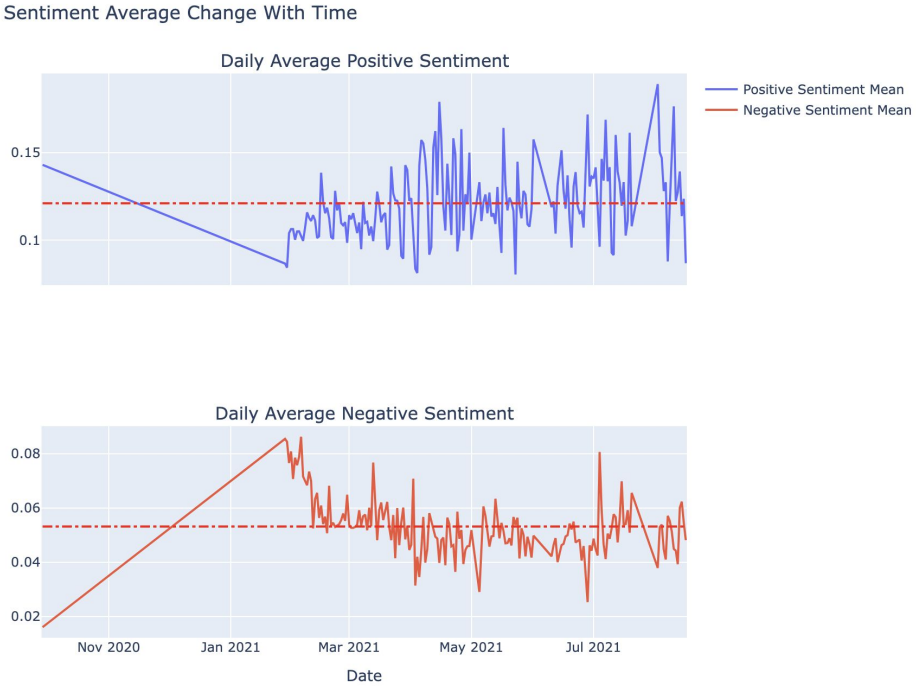
Disclaimer - The dataset is unfiltered, giving a sense of the colorful jargon used in the subreddit

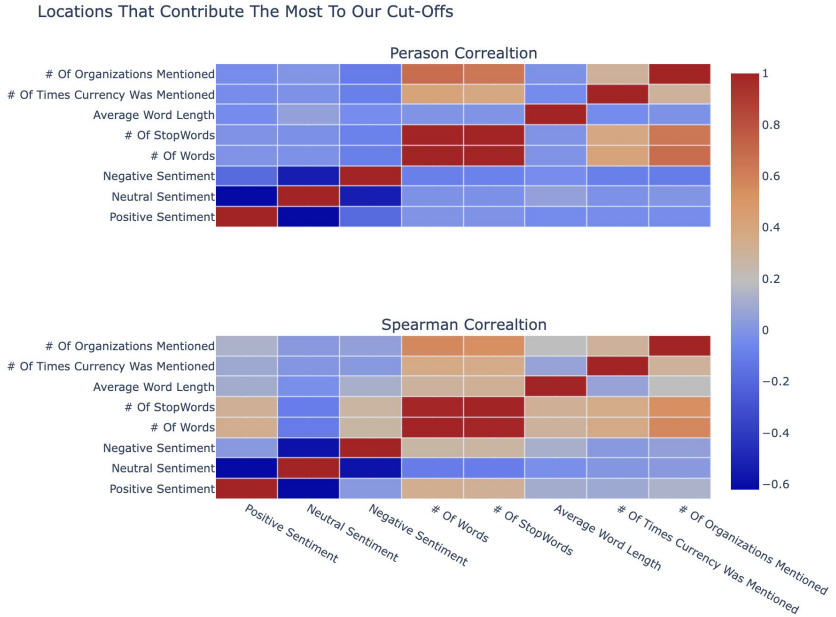
Analysis on Post Body text



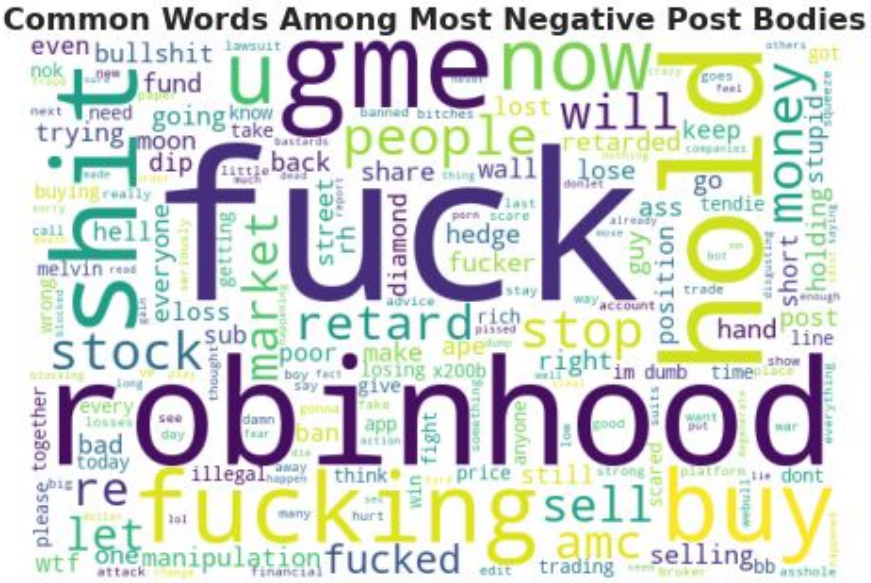


Analysis on Post Body text

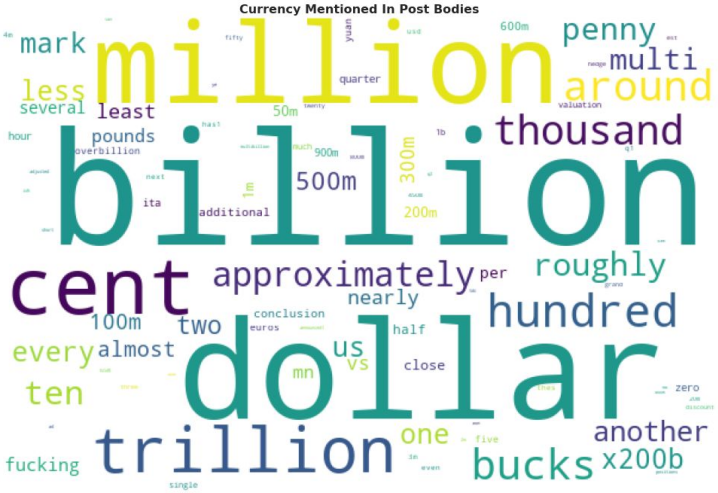




Analysis on Post Body text



Top Discussions in post



Sentiment Analysis Results

- Due to the visible trend both in the positive and negative sentiments of the post titles, we can conclude that the data is not stationary.

- There is a general increase in positive sentiment and a general decrease in negative sentiment over time.

- The majority sentiment of the subreddit has been tagged as neutral as seen in cdf distributions

- The average number of posts has remained constant throughout.

- The analysis model gives a very good brief about its correlation matrices and how different factors affect the sentiment of both the post titles and bodies.

- 82% of all the discussions in the subreddit can be done with 450 words.

- As of now the model lacks the capability to classify new statements but it sure can with a little tweaking in the code

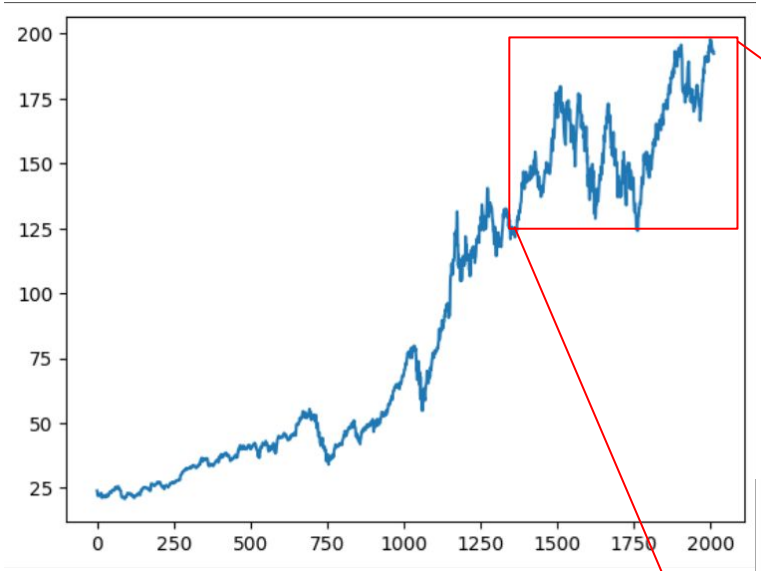
Proposed Architecture - Stock Predictor

We intend to use **Long Short Term Memory Neural Networks (LSTM)** for our architecture - majorly because of its ability to learn from sequential data while considering long- term dependencies. In the context of stock prediction, past stock prices and other relevant factors form a time series, and LSTM can effectively capture patterns and trends within this sequential data. Also their ability to handle non-linear data and memory retention, they seem to be the perfect model for our application.

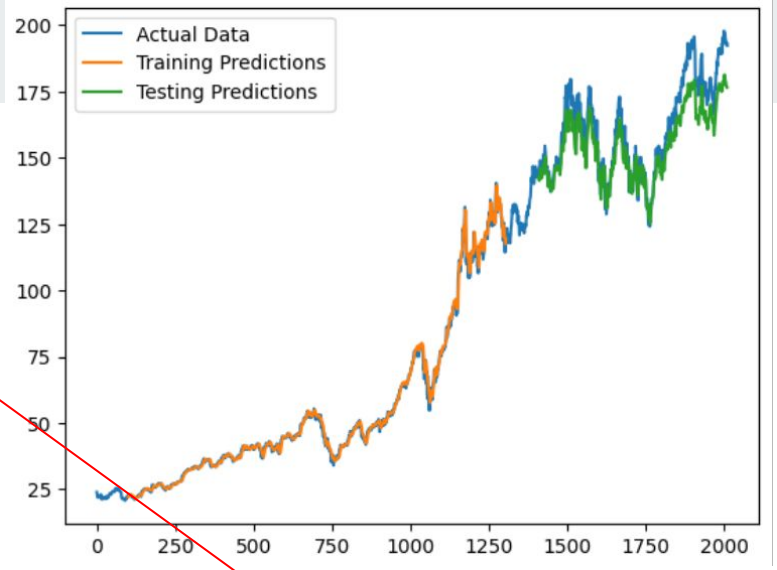
We have used Tensorflow's Sequential model with LSTM and Dense layers for libraries. The model trains and predicts for each day using the trends of the last 100 days to make its decisions. We have used 3 layers for LSTM and one layer for Dense, and have computed our losses in terms of mean- squared error.

In terms of dataset we have used the listings from January 2016 to January 2023, and have used 65% of the listings as train set and 35% as test set

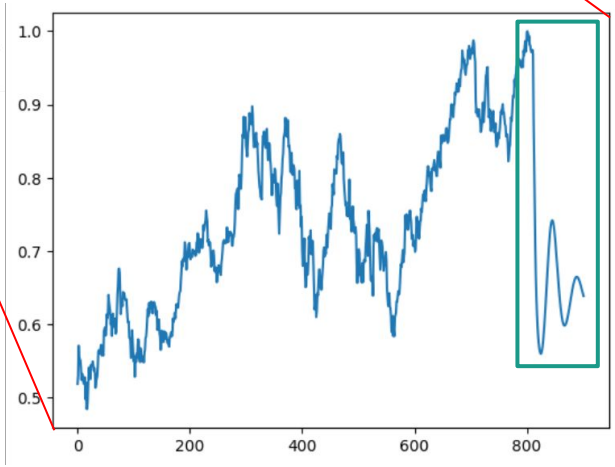
Stock Prediction on AAPL



Apple Stock - days to price

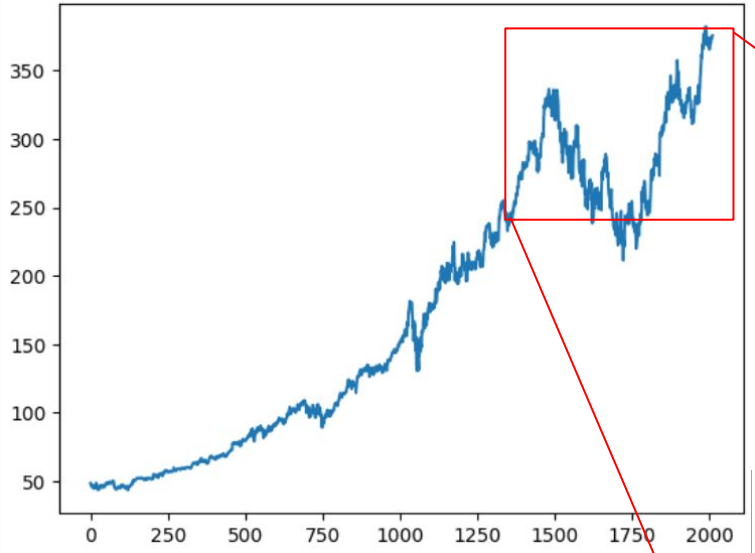


Train and Test predictions on the data

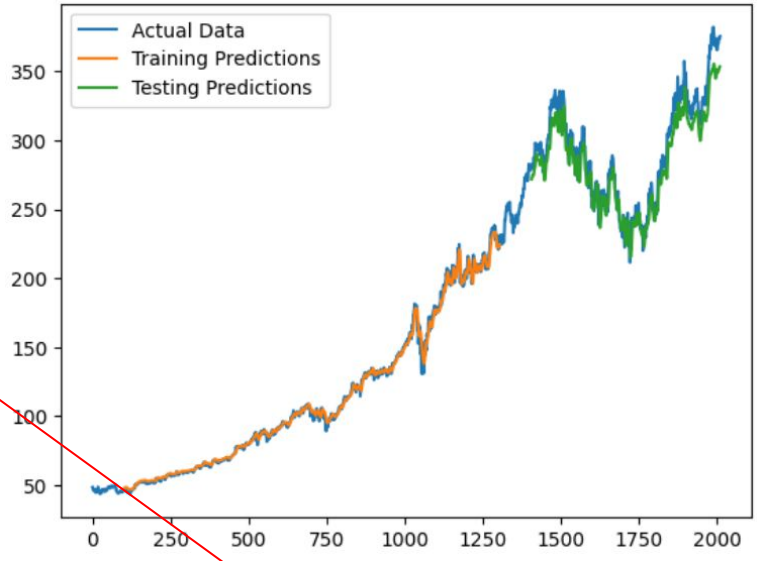


Apple predictions for the next 90 days

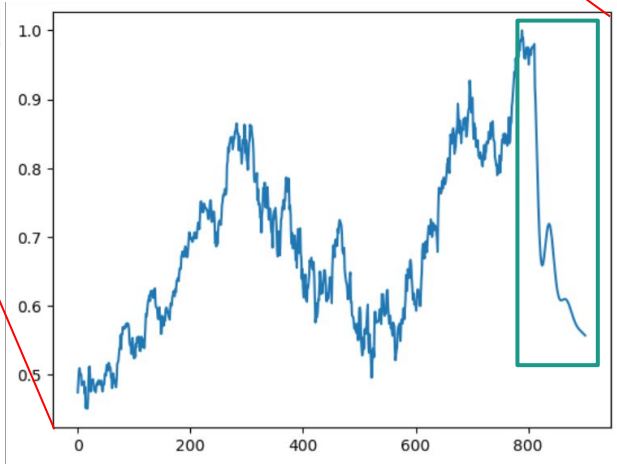
Stock Prediction on MSFT



Microsoft Stock - days to price



Train and Test predictions
on the data



Microsoft predictions for
the next 90 days

Stock Predictor Results

- LSTM proved to be a good regressor for something like stock prediction which is factored by a lot of real world variables which one might say to be random, but using the past data it was able to give a good estimate.
- The future predictions seem a little flawed considering the huge dips and rises in our predictions compared to the past trends.
- Our model fails to predict for large durations in the future because of its data being referenced from another future predicted data. At best, it can give a prediction for the next day or two.

Future Scope

- The sentiment analysis provides an in-depth analysis and understanding of the vocabulary and the jargon used in one of the world's biggest forums for stock trading, and thus can be leveraged to judge what the forum as a community decides upon to be a good decision or a bad decision in a field which is subjected to a lot of risk. It provides beginners a good starting point into trading.
- These results from sentiment analysis can be fact checked from a good stock predictor model to determine whether the strategies in the subreddit are viable or not
- But since the stock market is a very fragile system that is all subjected to nearly infinite real world factor - its very hard to come up with such a seemingly perfect predictor to give results based on only the past trends

References

- The project is deeply inspired by Michael Reeves video “I gave my goldfish $50,000 to trade stocks” which follows similar techniques of sentiment analysis
- Greg Hogg’s youtube video on LSTMs and Stock prediction model - https://www.youtube.com/watch?v=CbTU92pbDKw
- Krish Naik’s youtube video on LSTMs and Stock prediction model - https://www.youtube.com/watch?v=H6dupfuznE
- Sentiment Analysis Notebook on Kaggle based on subreddits by Thomas Konstantin - https://www.kaggle.com/code/accountstatus/wall-streets-bet-data-analysis?rvi=1
-

Thank you

Saksham Singh
Ritika Thakur