

S&P 500

Group Members

Beichi Wu

Qilin Zhou

Sizhe Ding

Yolanda Yang





Page Index

Group Members	1
Project Abstract	2
Dashboard Interaction Instruction	3
Description of Code Responsibilities	8
Description of Major Sections	9



Project Abstract

The market efficiency hypothesis posits that market prices encapsulate all relevant information. However, the overall societal mood with respect to a company, as expressed through news, is now viewed as a crucial variable that affects the stock price of the company. Rampant information and data have overwhelmed buyers and sellers, especially retail investors. Our project empowers retail investors by bridging the cognitive or knowledge gap in financial markets, augmenting investors' decision-making processes with advanced, data-driven insights.

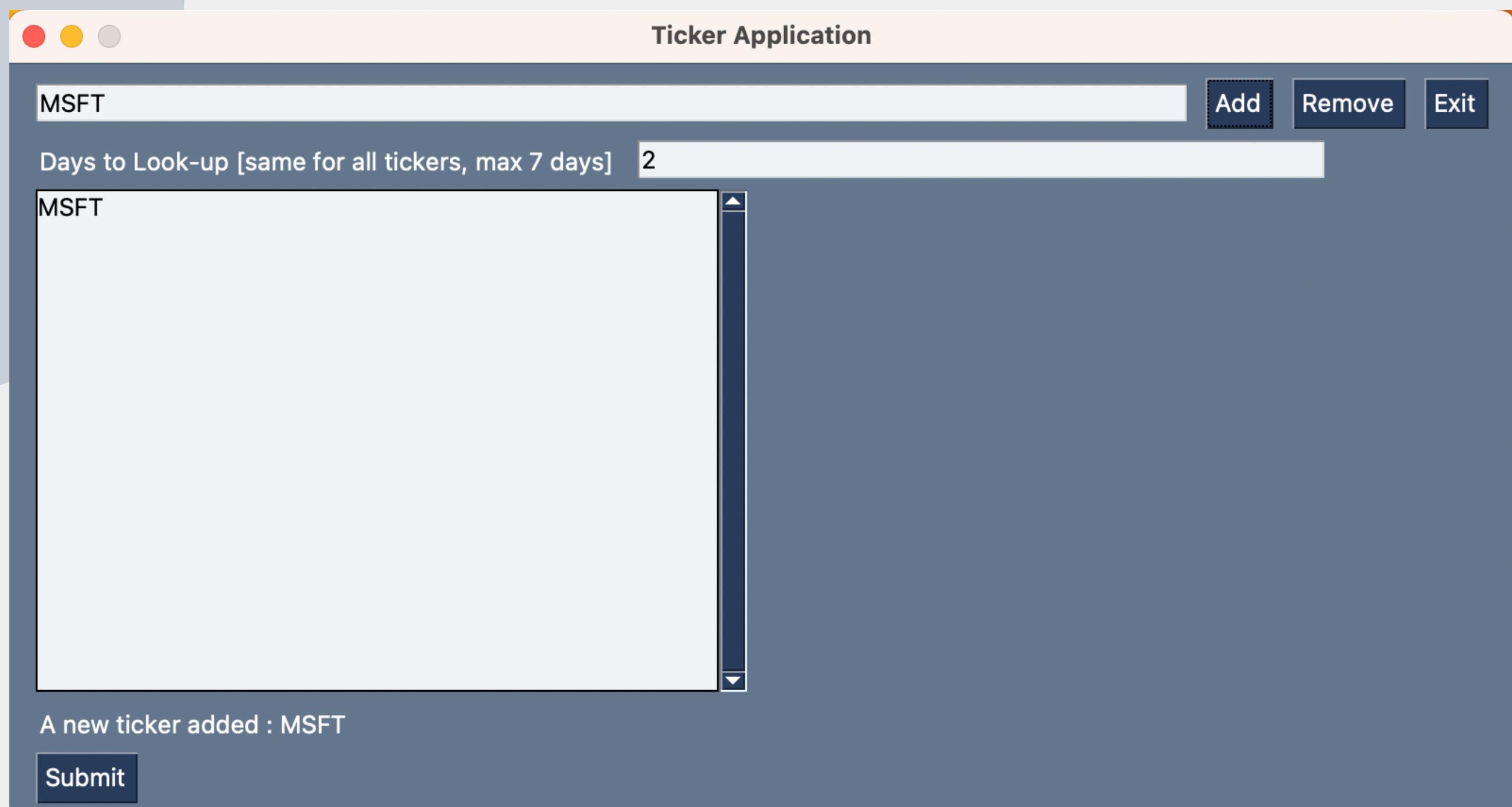
Our project, leveraging sentiment analysis, aims to capture the collective sentiment towards companies listed in the S&P 500 index by analyzing news from various sources. This approach seeks to address the challenge of information overload faced by retail investors. News articles related to the S&P 500 companies were sourced from the finviz.com website for this purpose. A word frequency-based analyzer was developed for sentiment analysis, complemented by manual sentiment labeling to generate training and test datasets.

Furthermore, the project incorporates time-lagged and macroeconomic factors into the analysis. Utilizing methods like random forests and artificial neural networks, we aim to enhance the visualization of stock movements and improve the accuracy of our predictive models. A review of existing literature reveals various time series prediction models such as ARIMA and Facebook Prophet, with findings indicating superior performance of Recurrent Neural Network (RNN) models in correlating textual information with stock price movements. Our research expands upon this foundation by employing three distinct time series models: Artificial Neural Networks, Long Short-Term Memory networks, and Random Forest models.

Interacting with Dashboard

The dashboard includes 5 panels for user interaction:

1. Input tickers: the dashboard allows unlimited number of company tickers that are within Yahoo Finance and up to 10 days
 - This shows a list of news headlines for the companies for the duration of days that you inquired, you can click on the headlines to read more on each article

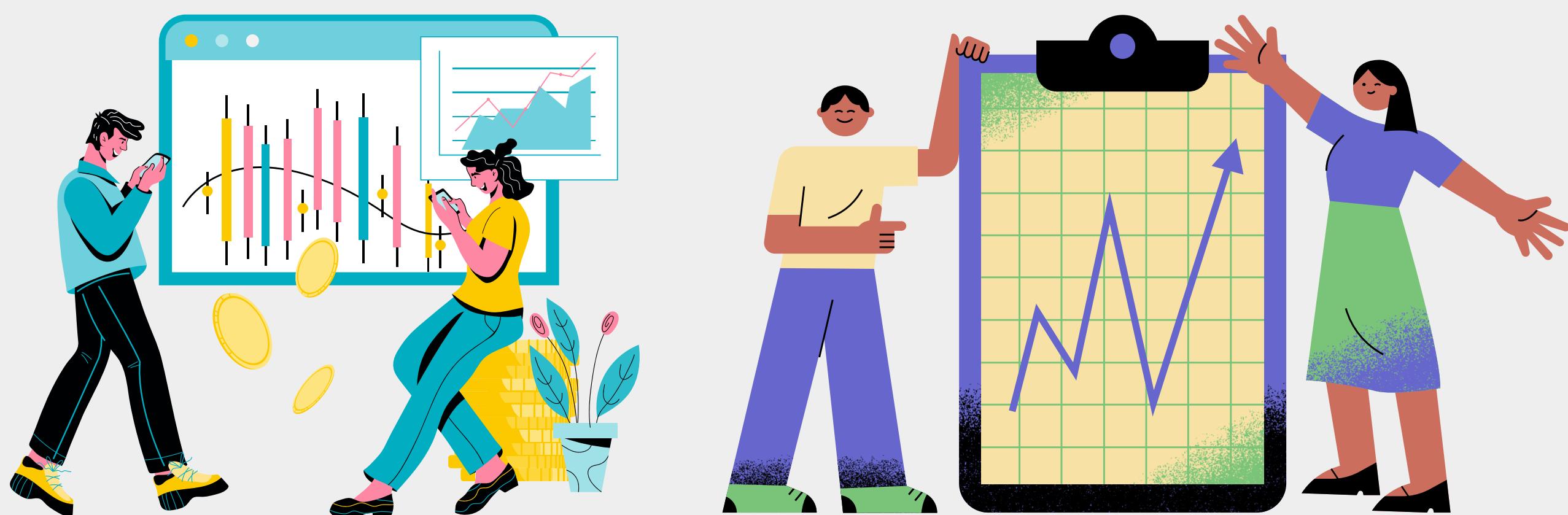
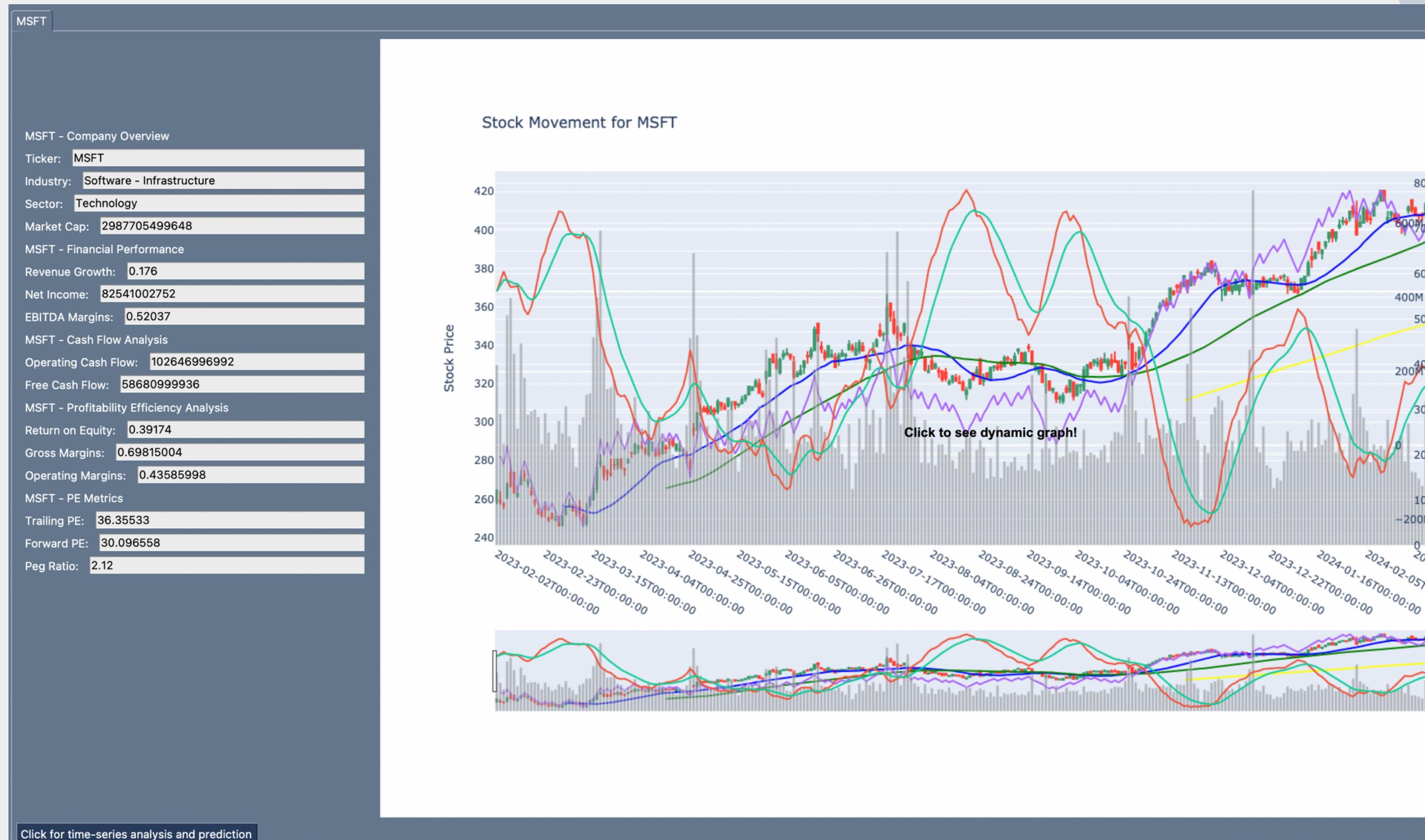


Overview with URLs Clickable					
Date	Time	Company	Headline	URL	
2024-03-07	06:33AM	MSFT	UPDATE 1-As Big Tech scrambles to meet EU rules, investigations seen as likely	https://finance	
2024-03-07	05:40AM	MSFT	This 1 Chart Shows How Microsoft Became the Biggest "Magnificent Seven" Stock and World's Mo	https://finance	
2024-03-07	05:06AM	MSFT	Artificial Intelligence (AI) Stock Nvidia May Be the Bubble of the Century, and History Suggests It W	https://finance	
2024-03-07	04:56AM	MSFT	OpenAIs success has put a huge target on its back	https://finance	
2024-03-07	01:03AM	MSFT	As Big Tech scrambles to meet EU rules, investigations seen as likely	https://finance	
2024-03-06	06:01PM	MSFT	Top AI photo generators produce misleading election-related images, study finds	https://finance	
2024-03-06	05:45PM	MSFT	Microsoft (MSFT) Stock Slides as Market Rises: Facts to Know Before You Trade	https://finance	
2024-03-06	04:37PM	MSFT	Magnificent Seven Stocks To Buy And Watch: Nvidia Hits Record High, Tesla Slides	https://finance	
2024-03-06	04:36PM	MSFT	UPDATE 3-LinkedIn back up following brief outage	https://finance	
2024-03-06	04:26PM	MSFT	Box Stock Jumps On Earnings Beat, Microsoft AI Collaboration News	https://finance	
2024-03-06	04:23PM	MSFT	LinkedIn back up following brief outage	https://finance	
2024-03-06	04:21PM	MSFT	LinkedIn down for thousands of users, Downdetector shows	https://finance	
2024-03-06	04:06PM	MSFT	Imagine Elon Musk as Donald Trumps boss	https://finance	
2024-03-06	03:59PM	MSFT	Microsoft's engineer warns company's AI tool creates problematic images	https://www.yc	
2024-03-06	03:31PM	MSFT	Microsofts AI Tool Generates Sexually Harmful and Violent Images, Engineer Warns	https://finance	
2024-03-06	03:20PM	MSFT	CrowdStrike Scores With Fourth Quarter Results, Outlook And Market Expanding Acquisition	https://finance	
2024-03-06	03:17PM	MSFT	30 Most Educated Cities In The World	https://www.in	
2024-03-06	03:13PM	MSFT	ChatGPT: Everything you need to know about the AI-powered chatbot	https://finance	
2024-03-06	02:54PM	MSFT	OpenAI publishes Elon Musks emails. Were sad that its come to this	https://finance	
2024-03-06	02:48PM	MSFT	Microsoft engineer sounds alarm on AI image-generator to US officials and company's board	https://finance	
2024-03-06	02:06PM	MSFT	Microsoft employee: AI tool should be removed until offensive images can be addressed	https://finance	
2024-03-06	01:19PM	MSFT	It's not just Elon Musk: ChatGPT-maker OpenAI confronting a mountain of legal challenges	https://finance	
2024-03-06	12:57PM	MSFT	OpenAI says Musk only ever contributed \$45M, wanted to merge with Tesla or take control	https://finance	
2024-03-06	12:17PM	MSFT	Top Companies for Employee Engagement and Development	https://finance	
2024-03-06	11:55AM	MSFT	Microsoft engineer says company was 'aware of the potential for abuse' before viral Taylor Swift de	https://finance	
2024-03-06	11:46AM	MSFT	A Microsoft employee warns that the companys AI tool can generate offensive images	https://finance	
2024-03-06	10:52AM	MSFT	OpenAI and Microsoft's AI bot Copilot are on the hot seat for sexual and copyright-violating images	https://qz.com	
2024-03-06	10:32AM	MSFT	Billionaire Jeff Bezos and Cathie Wood Joined Microsoft, Nvidia, Amazon, and OpenAI in Funding 1	https://finance	
2024-03-06	09:15AM	MSFT	1 Brilliant Strategy That Could Supercharge Microsoft's Artificial Intelligence (AI) Revenue	https://finance	
2024-03-06	08:34AM	MSFT	Q4 2024 CrowdStrike Holdings Inc Earnings Call	https://finance	
2024-03-06	08:31AM	MSFT	SOPHIA GENETICS SA (NASDAQ:SOPH) Q4 2023 Earnings Call Transcript	https://www.in	
2024-03-06	08:04AM	MSFT	OpenAI Hits Back at Elon Musk's Lawsuit, Alleging Billionaire Was Also Behind Profit Push	https://www.in	
2024-03-06	07:28AM	MSFT	5 Things to Know Before the Stock Market Opens	https://www.in	
2024-03-06	07:20AM	MSFT	Best Stocks To Buy And Watch Now: Palantir Headlines 5 Top Tech Stocks For March	https://finance	
2024-03-06	06:50AM	MSFT	Stocks, Gold, and Bitcoin Pull Back From Record Highs, What It Means. And 5 Other Things to Know	https://finance	
2024-03-06	06:41AM	MSFT	Elon Musk Wanted OpenAI to Merge With Tesla, Altman Says	https://finance	
2024-03-06	06:16AM	MSFT	Should You Buy the Worst-Performing "Magnificent Seven" Stock of the Past Decade?	https://finance	
2024-03-06	06:15AM	MSFT	Nvidia, Microsoft, and Jeff Bezos Invested in a \$2.6 Billion Robotics Start-Up. Here's Why That's G	https://finance	
2024-03-06	06:10AM	MSFT	Still Magnificent: Why MSFT Stock Is Headed to \$475 (or Higher)	https://investo	
2024-03-06	06:05AM	MSFT	OpenAI, Microsoft AI tools generate misleading election images, researchers say	https://finance	
2024-03-06	05:25AM	MSFT	Microsoft's Path to Becoming the Largest Company in the World, Explained in One Chart.	https://finance	
2024-03-06	05:14AM	MSFT	Microsoft launches new initiatives to skill and scale AI transformation in Singapore	https://finance	
2024-03-06	05:06AM	MSFT	1 "Magnificent Seven" Stock to Buy Hand Over Fist in March, and 1 to Avoid Like the Plague	https://finance	
2024-03-06	04:02AM	MSFT	OpenAI fires back at Musk allegations with trove of emails	https://www.di	
2024-03-05	10:30PM	MSFT	UPDATE 1-OpenAI seeks to dismiss all of Musk's claims in lawsuit	https://finance	

For More Detailed Analysis
(First Press)Sentiment Analysis
Sample Word Clouds
(Second Press)Choose One Company for Word Clouds

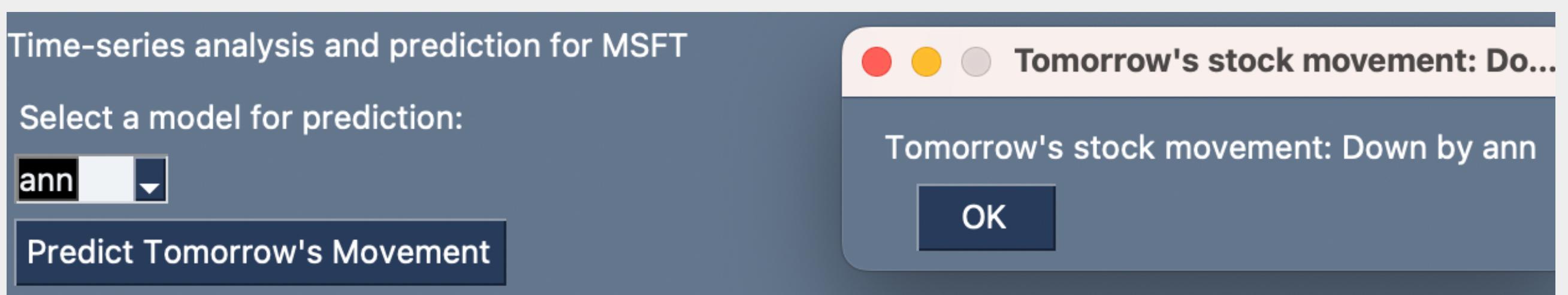
2.4. For more detailed analysis: interactive stock movements

- you can view company overview, and click to view each company's stock movements with different indicators such as price details, volume, MACD, etc.



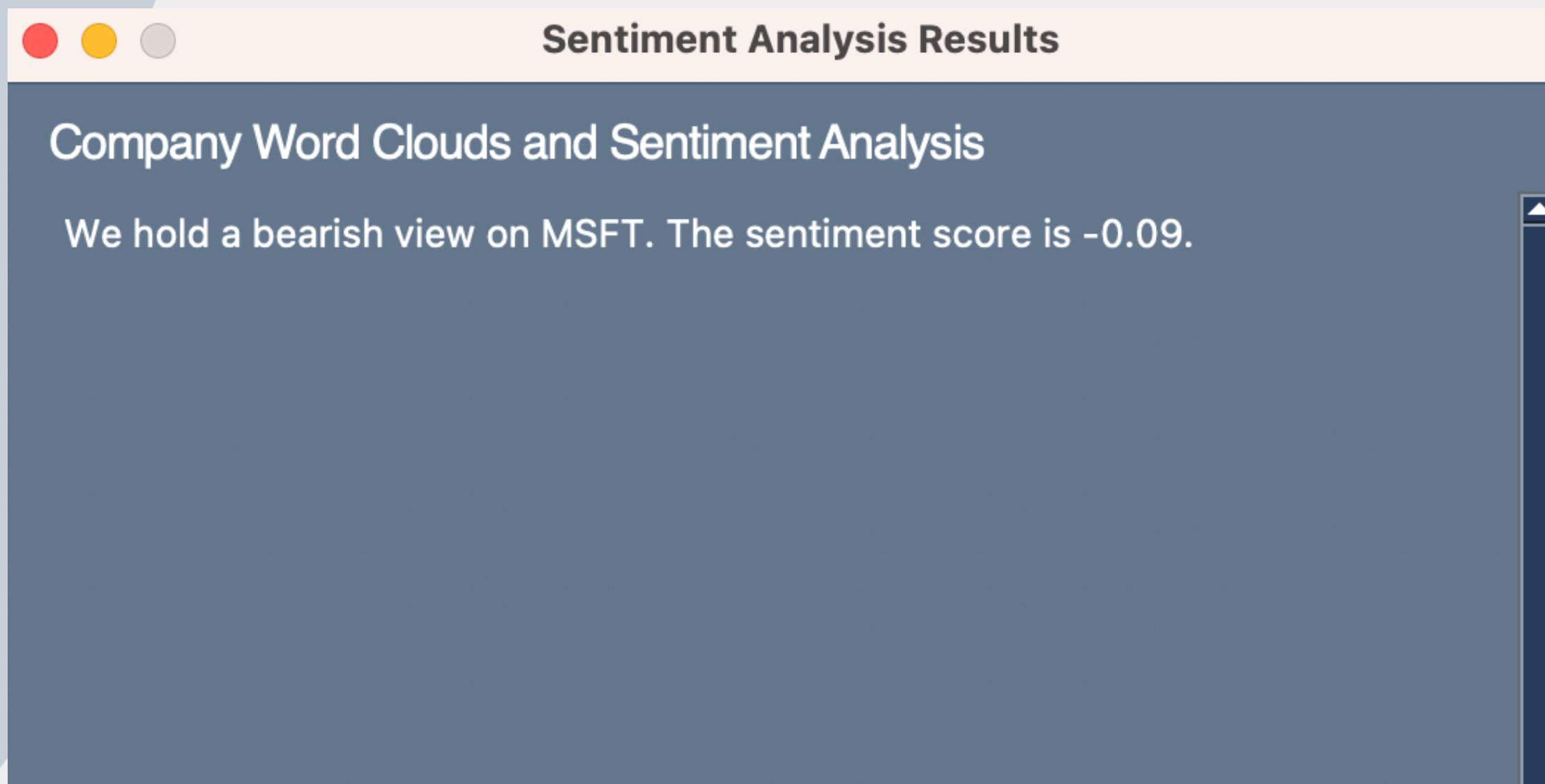
2.1. Time Series Models

- you can select from Artificial Neural Network, Long Short Term Memory, and Random Forest models to view the corresponding model's prediction of the stock price



3. First Press: Sentiment analysis:

- you can view the sentiment score calculated based on classifier



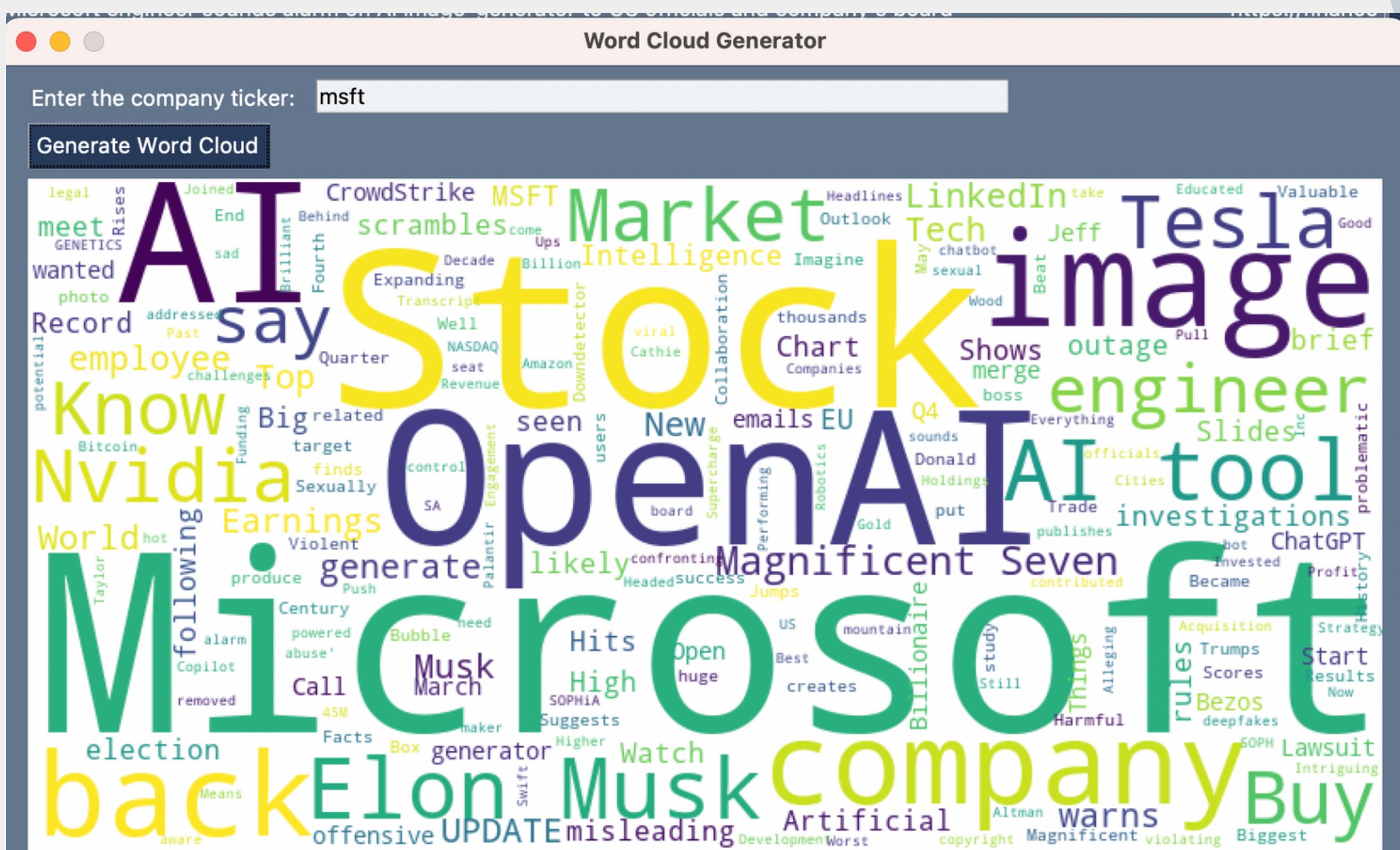
4. Sample Word Cloud picture:

- after concluding the first press, you can view a sample word cloud if the inquiry company is one of the following five: AAPL, GOOG, NVDA, AMZN, and BA



5. Second Press: Choose One Company for Word Cloud:

- you can input a company ticker to view that company's word cloud



Description of Code Responsibilities

	Responsibility Summary	Associated Files/packages
Qilin Zhou	Research: Understanding APIs, researching different machine learning models, and scrape text	
	Data Collection	<ul style="list-style-type: none"> - headlines - scraper.py
	Graphic User Interface Design	<ul style="list-style-type: none"> - headlines - app.py - displayer.py
	Time-Series Analysis	<ul style="list-style-type: none"> - time_series - price_model.py - time_series_preprocessing.py
	Document writing	README.md
Beichi Wu	Research: Researching methods of sentiment analysis and design a supervised way of training dataset; Learning quantitative finance	
	Data Cleanup, Compile, Part of Speech Tagging	<ul style="list-style-type: none"> - compile - cleanup.py - pos.py
	Sentiment Analysis	<ul style="list-style-type: none"> - sa - analyzer.py - sa.py - test.py - train_classifier.py
	Document writing, Project folder design	README.md Project Paper.pdf
Sizhe Ding	Research: researching methods of how to use different colors in word clouds to express various emotions, aiming to achieve the visualization of sentiment analysis.	
	Visualization	<ul style="list-style-type: none"> - Visualization - create_word_cloud.py - datatypes.py - wordcloud_in_logo.py
Yolanda Yang	Research: Understanding different machine learning models, learning the process of quantitative finance	
	Sentiment Analysis: Data Labeling	- Jan_24_Jan_28_Stock_News.csv
	Time-series Analysis and Visualization	<ul style="list-style-type: none"> - time_series - visualization - macro_indicators_viz.py - stock_movement.py
	Document Writing	Project Paper.pdf

Each member of the team encountered numerous difficulties while completing their respective tasks, yet everyone excelled in completing the assignment!

Description of major sections

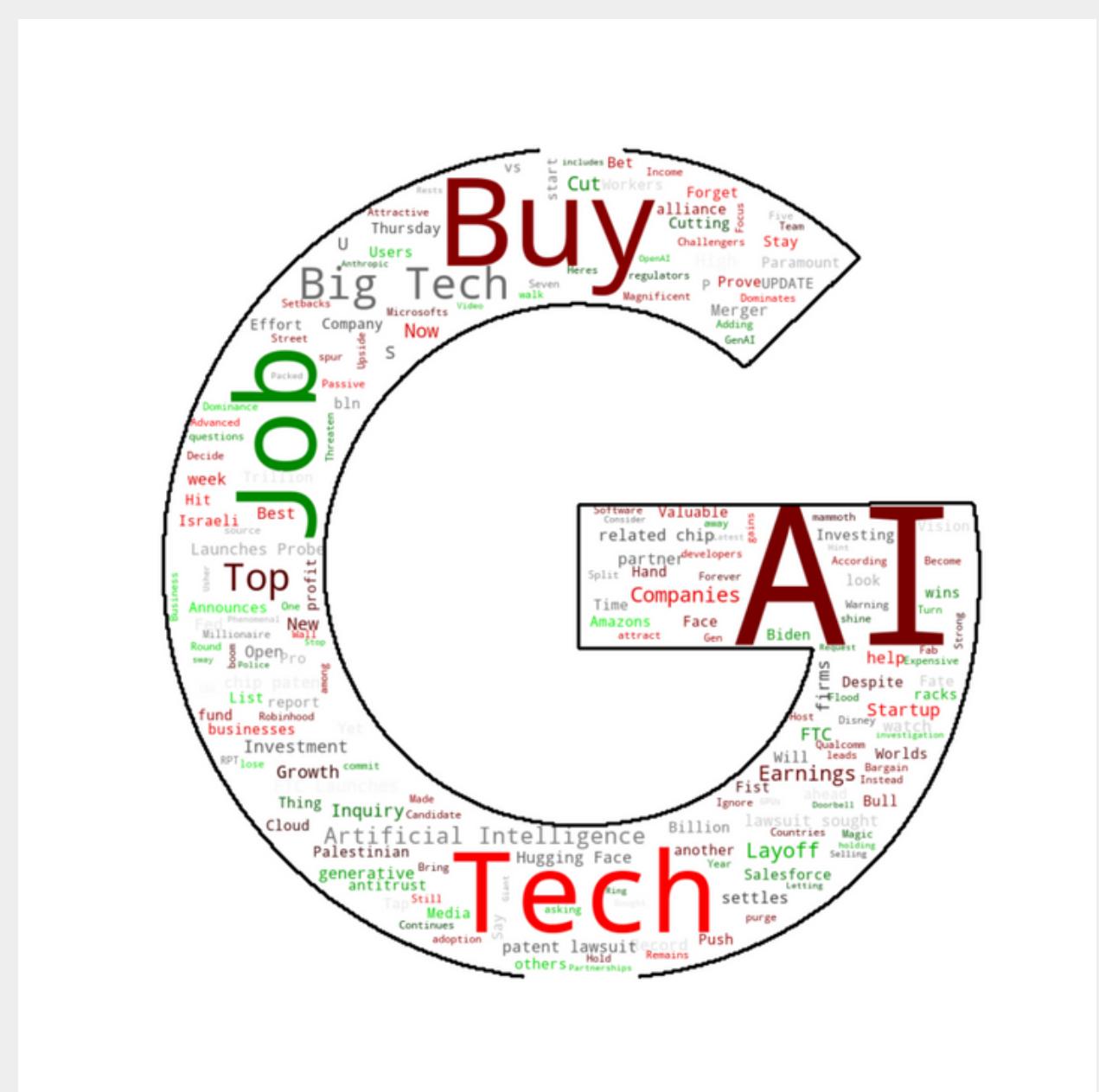
Sentiment Analysis (Sentiment Score, Classifier Training & Word Cloud)

I. Financial news headlines reflect investors' sentiment. In this project, we aim to web scrape financial news headlines from finviz.com and analyze their sentiment accordingly.

- a. Firstly, we scraped the news headlines for Apple, Nvidia, Boeing, Amazon, and Google from January 24 to January 28. We manually labeled over 400 news headlines to determine the sentiment ("P", "N", "U") of the news headlines.
 - b. Then, based on the labeled dataset, we trained a classifier. We applied the nltk and textblob packages to calculate sentiment scores, ranging from -1 to 1, for the news headlines. To obtain more precise polarity scores, we fine-tuned the nltk package using the Loughran-McDonald Master Dictionary, which includes words appearing in 10-K documents and earnings calls..
 - c. We further divided the sentiment scores into five categories: "Positive High", "Positive Low", "Negative Low", "Negative High", and "Neutral". There are 25 possible tuples ranging from ("Negative High", "Negative High") to ("Positive High", "Positive High"). These tuples were mapped to our manual labels.
 - d. We calculated the possibilities, $m(t)$, the chance of seeing one label according to one tuple. The mapping process was recorded as our classifier.
 - e. For newly inputted dataframes, the algorithm calculates the nltk and textblob score tuples based on the news headlines and determines the final sentiment by using the classifier obtained above.

II. Use the wordcloud package to generate word clouds in the shape of each company's logo for every company, and recolor the word clouds.

- a. Red represents words analyzed as positive, green represents words analyzed as negative, and gray represents neutral or indeterminate words.
 - b. The size of the font represents the frequency of word occurrence. By observing the overall color distribution of the cloud and the prominent words, one can briefly judge the attitude of the news towards a certain stock during this period, facilitating the combination with subsequent sentiment analysis.



Time series Analysis

- I. Utilize DBnomics and FRED APIs to scrape and process macro-economic indicators such as interest rates and unemployment rates. Gather historical stock data using yfinance library and calculate indicators such as SMA, EMA, MACD, and OVB. Train the time-series model with data from 2000-01-01 to the current date with optional backtesting validation, combine stock data with macroeconomic indicators, and apply preprocessing steps for the stock movement prediction model.
- II. Previous literature reveals that creating a hybrid model using machine learning models such as LSTM and sentiment analysis model based on news data may provide more accurate predictions of long and short-term stock price movements. Mohan and Mullapudi suggest that models that take financial news as part of the input have performed very well, while those predicted stock prices merely on historical stock prices lead to high percentage error.
- a. Train and evaluate three machine learning models for stock price movements: ANN, RNF, LSTM;
 - b. Conduct hyperparameter tuning for model configuration, evaluate performance on test data and backtest the model with five-year as a cycle;
 - c. Visualize the model predictions by creating a confusion matrix heatmap and probability plot for both positive and negative class probabilities, generate prediction summary plots and accuracy metrics for each model.

Training Period Performance for AAPL (2000-01-01 to 2024-03-03)

Model	Accuracy	Precision	Test Positive Percentage	Recall	F1 Score
RF	0.5554	0.5721	0.5499	0.7605	0.6529
LSTM	0.5421	0.5421	0.5421	1.0000	0.7031
ANN	0.5510	0.5505	0.5499	1.0000	0.7101

(note: Test Positive Percentage means the percentage of positive cases in the original data)

Back Testing Performance

Model	Accuracy	Precision	Test Positive Percentage	Recall	F1 Score
RF	0.4901	0.5370	0.5279	0.2484	0.3397
LSTM	0.5385	0.5385	0.5385	1.0000	0.7000
ANN	0.5279	0.5280	0.5279	0.9989	0.6908

We have **slightly better prediction than baseline distribution** (i.e., comparing precision with test positive percentage)