S&P 500

Group Members

Beichi Wu
Qilin Zhou
Sizhe Ding
Yolanda Yang

Page Index

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.

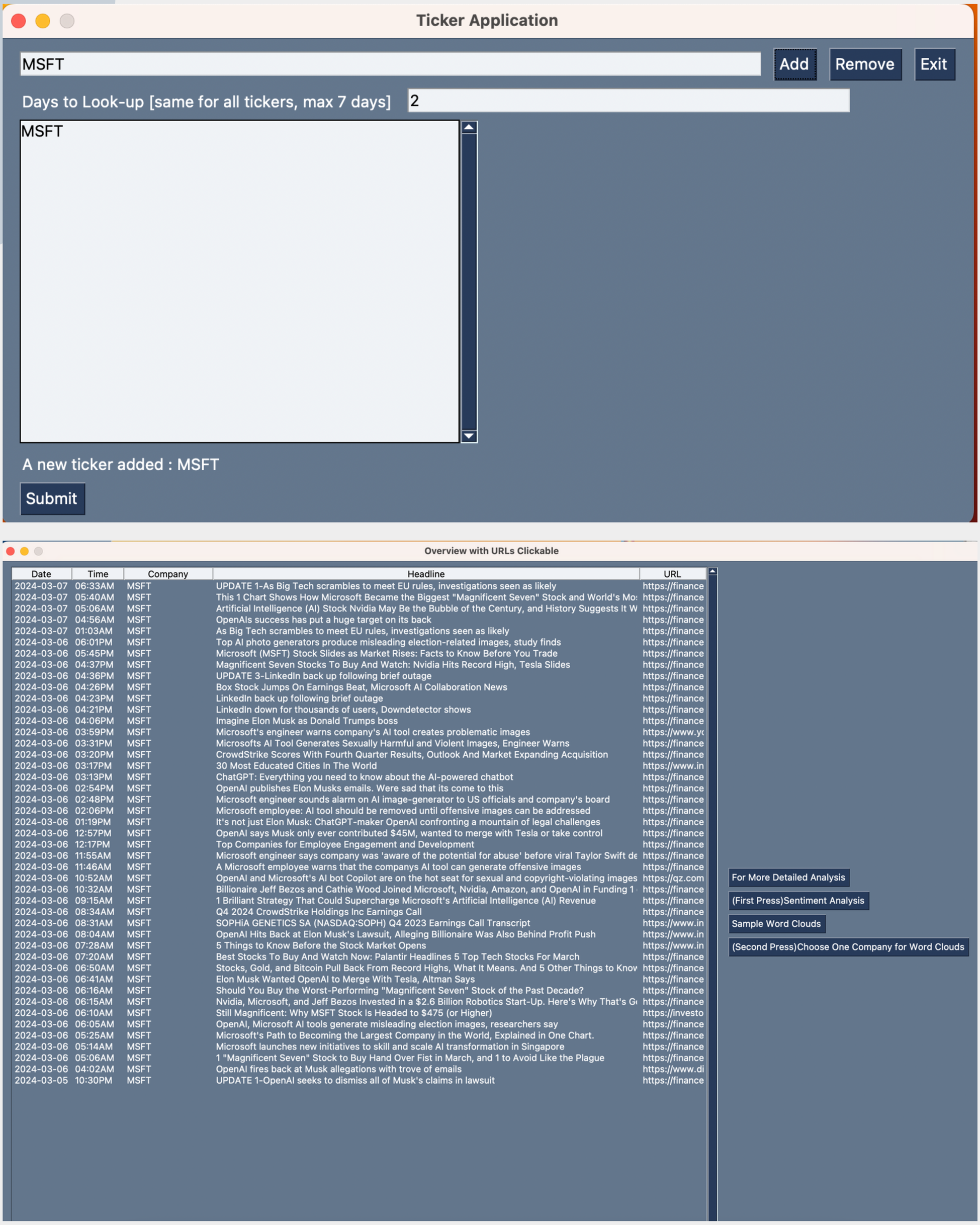
Our project, leveraging sentiment analysis, aims to capture the collective sentiment towards companies listed in the S&P 500 index by analyzing news from various sources. This approach seeks to address the challenge of information overload faced by retail investors. News articles related to the S&P 500 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

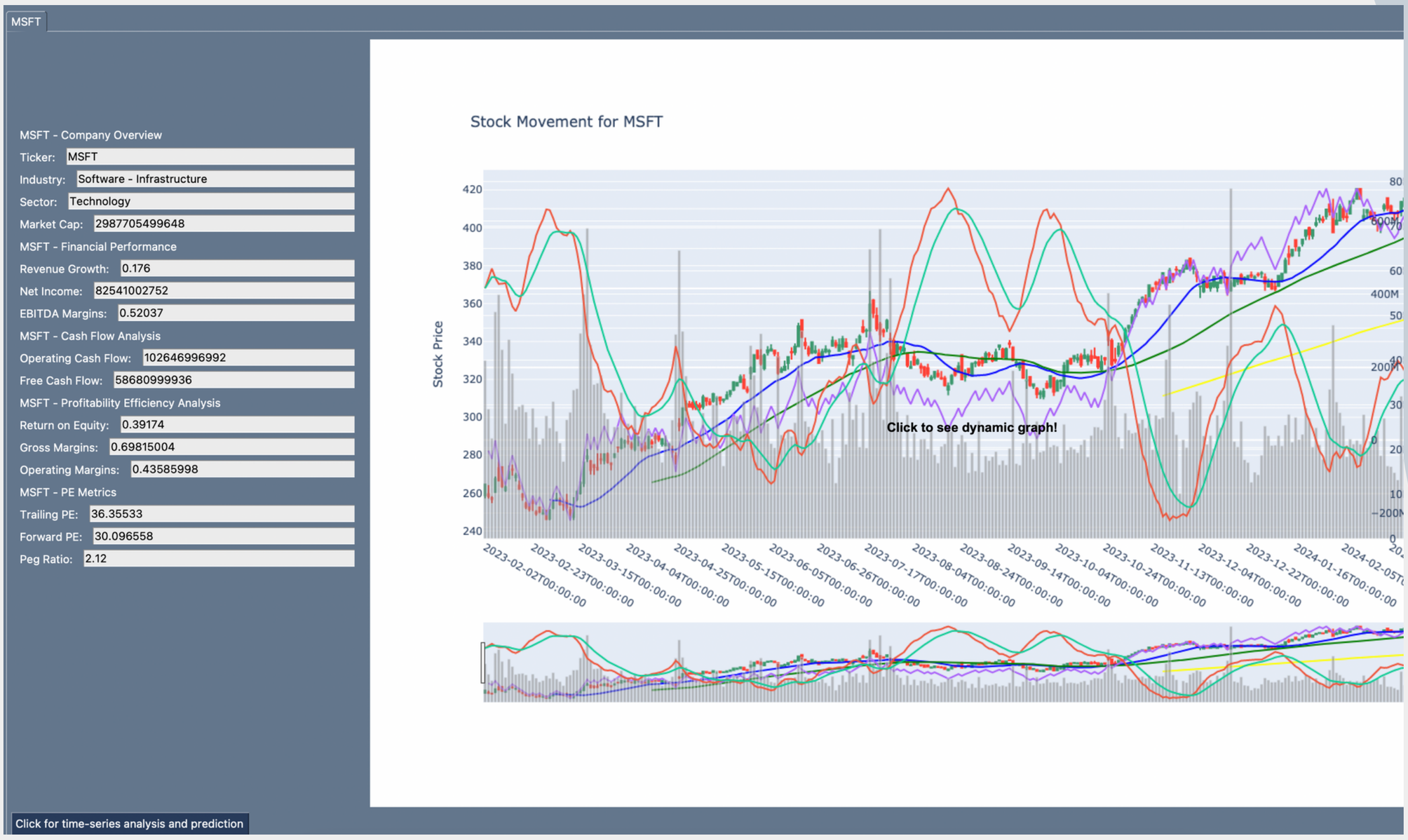
The dashboard includes 5 panels for user interaction:

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



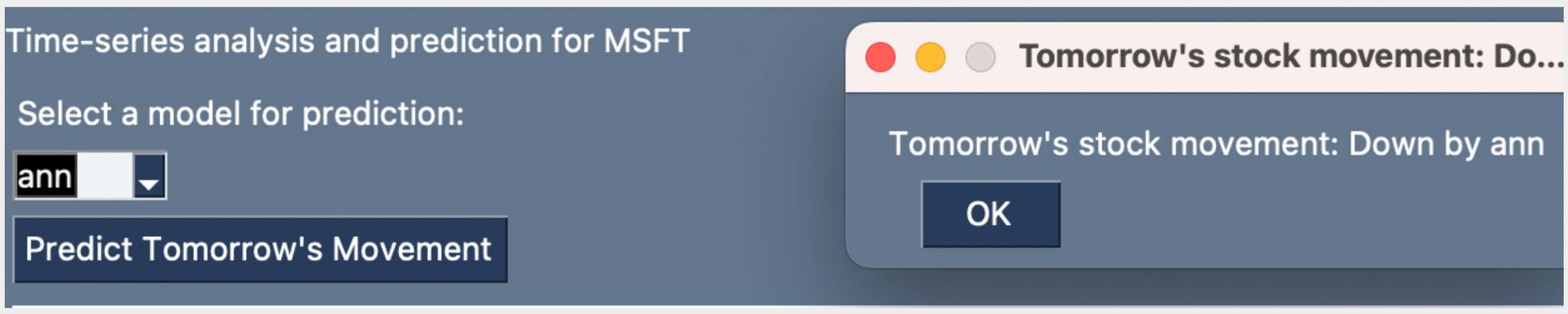
2.1. For more detailed analysis: interactive stock movements

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

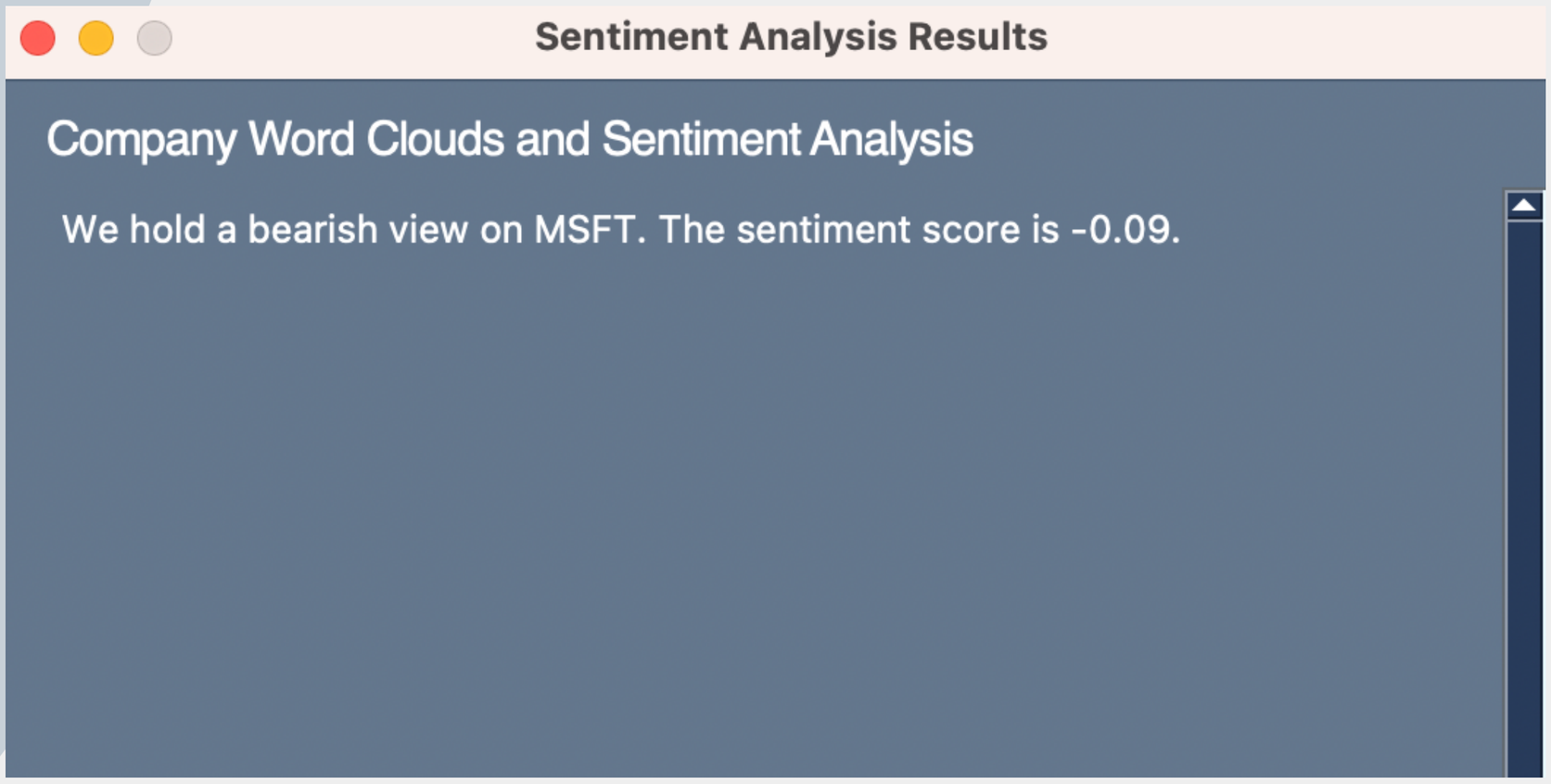


2.1. Time Series Model:s

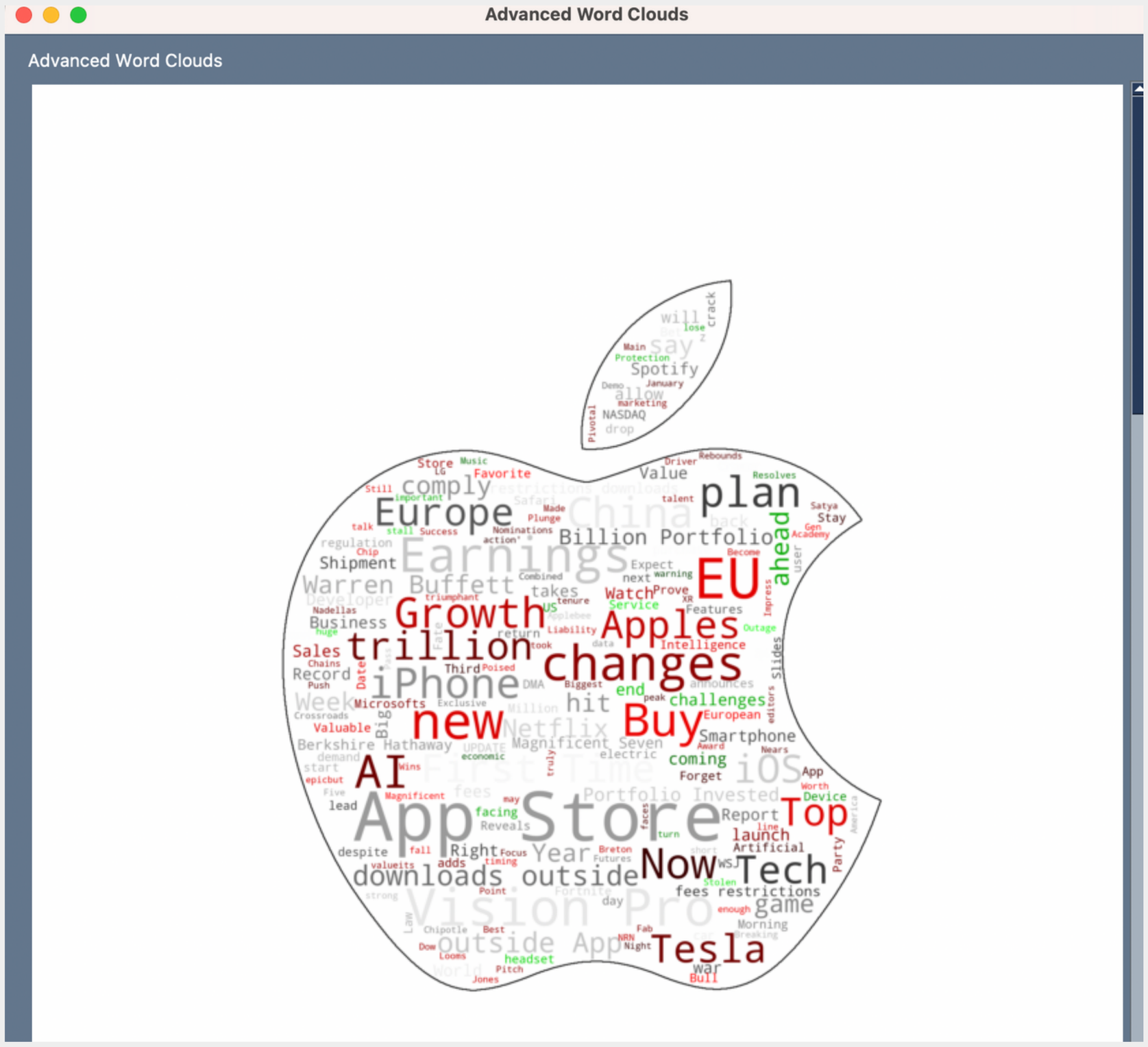
- you can select from Artificial Neural Network, Long Short Term Memory, and Random Forest models to view the corresponding model's prediction of the stock price



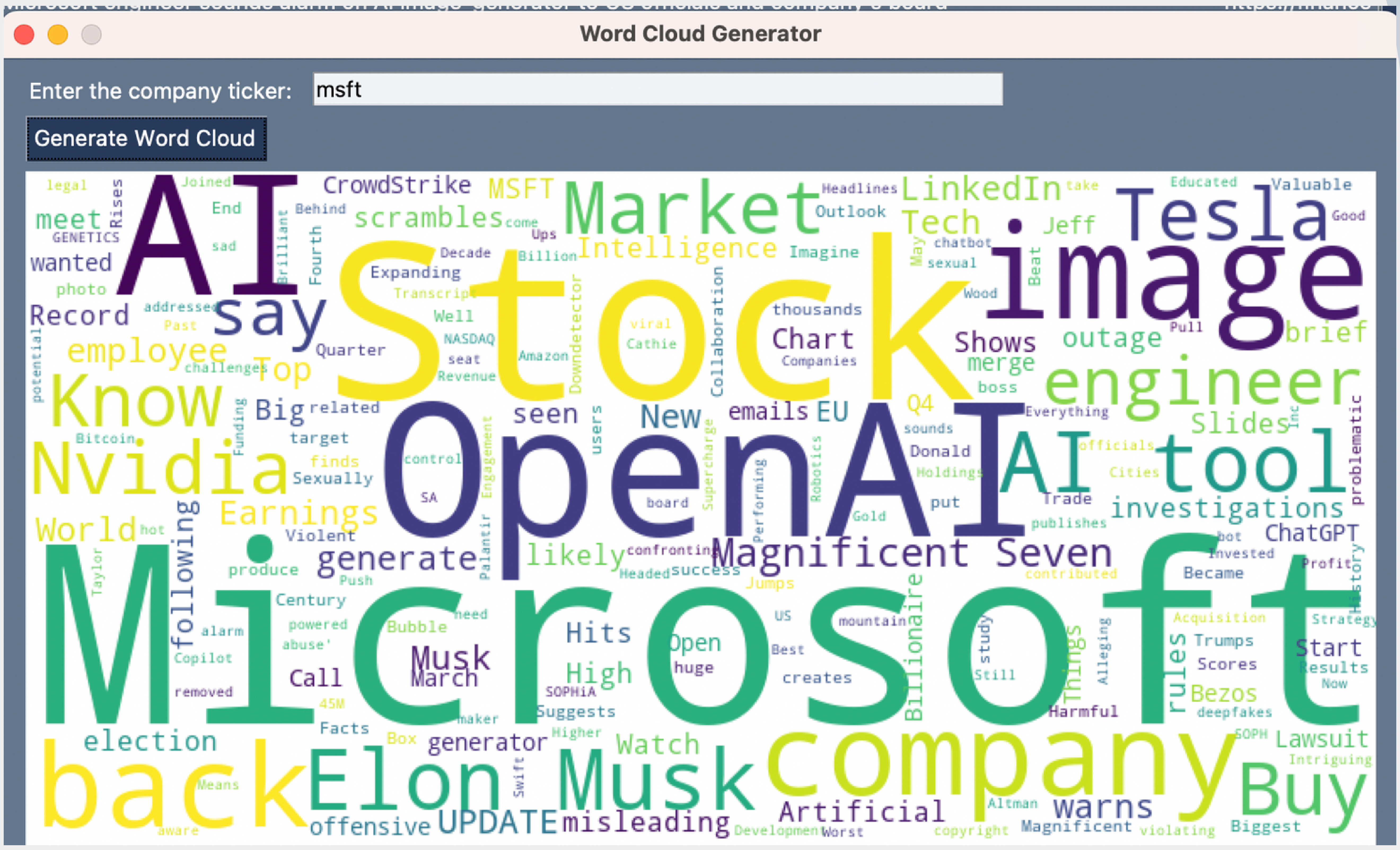
3. First Press: Sentiment analysis:
    - you can view the sentiment score calculated based on classifier



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	– headlines – scraper.py
	Graphic User Interface Design	– headlines – app.py – displayer.py
	Time-Series Analysis	– time_series – price_model.py – time_series_preprocessing.py – time_series – visualization – best_model_viz.py – company_profile.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	– compile – cleanup.py – pos.py
	Sentiment Analysis	– 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	–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	– 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

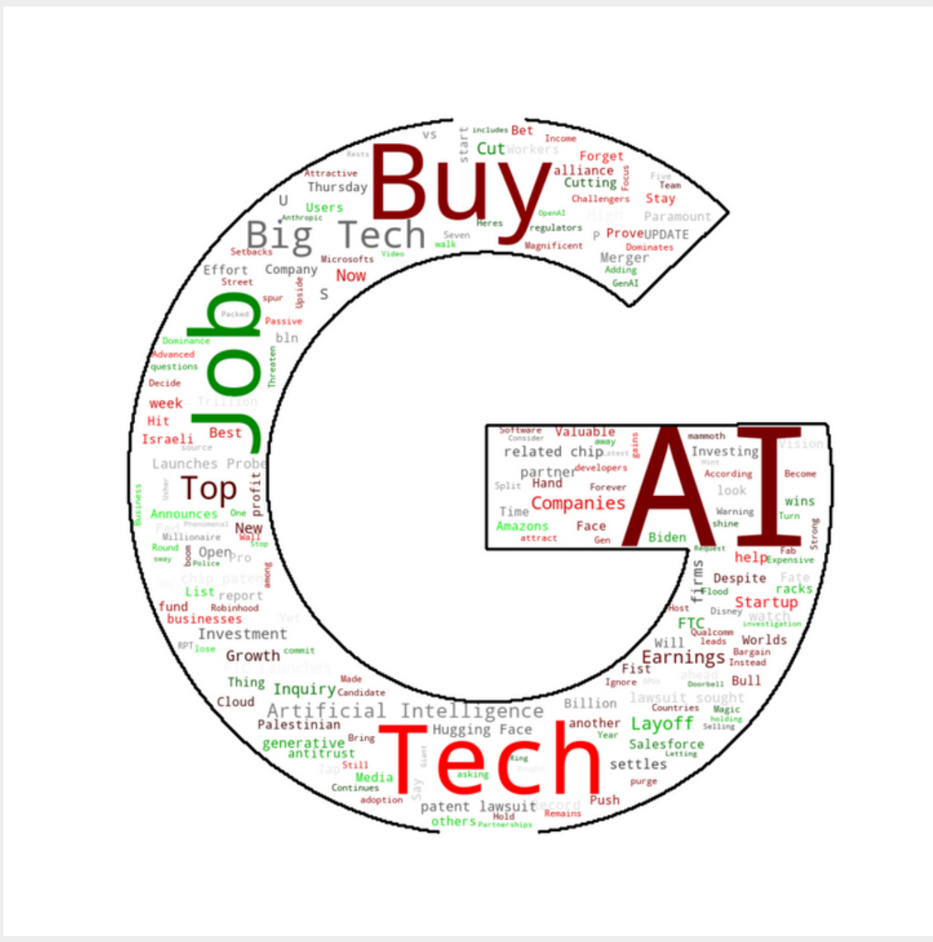
Sentiment Analysis (Sentiment Score, Classifier Training & Word Cloud)

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)

Beck 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)