UConn Team

1. Introduction:

The research project focuses on implementing sentiment analysis for stock data to gain valuable insights into market sentiment and its potential impact on stock prices. By analyzing news articles related to specific stock tickers, this project aims to provide investors and traders with a comprehensive understanding of market sentiment, allowing for more informed decision-making and improved investment strategies.

2. Goal:

The primary goal of the research project is to develop a Python script that fetches stock data from Yahoo Finance and retrieves relevant news articles using the Yahoo Finance API based upon a ticker. Subsequently, the project will perform sentiment analysis on these articles to determine the sentiment expressed in them. The specific objectives include:

- Retrieving stock data (e.g., opening price, closing price, and trading volume) for a given stock ticker from Yahoo Finance.
- Scraping news articles related to the selected stock to gather valuable textual data.
- Implementing sentiment analysis on the collected news articles to categorize sentiment as positive, negative, or neutral.
- Providing valuable insights into market sentiment and its potential impact on the stock price.

Relevant Definitions

In this research project, several crucial terms and concepts are used. Here are their accurate and detailed definitions:

- Sentiment Analysis: Also known as opinion mining, sentiment analysis is a Natural Language Processing (NLP) technique that involves the use of computational algorithms to identify, extract, and quantify subjective information from textual data. The goal is to determine the sentiment or emotion expressed in the text, which can be positive, negative, or neutral.

- NLP (Natural Language Processing): NLP is a branch of artificial intelligence that focuses on the interaction between humans and computers using natural language. It involves the development of algorithms and models to understand and process human language, enabling machines to interpret and respond to textual data.

- Web Scraping: Web scraping is the process of automatically extracting data from websites. In this project, web scraping is used to collect news articles related to the selected stock from reputable financial news sources on the internet.

- Yahoo Finance API: An Application Programming Interface (API) provided by Yahoo Finance, enabling developers to access financial data, including historical stock prices, trading volumes, and other relevant metrics. This API is utilized to retrieve stock data for the specified stock ticker.

- Lexicon-based Sentiment Analysis: A sentiment analysis approach that involves using sentiment lexicons or dictionaries containing pre-assigned sentiment scores to words. The sentiment scores of words in the text are aggregated to calculate an overall sentiment score for the document.

- Machine Learning-based Sentiment Analysis: An alternative approach to sentiment analysis, where machine learning algorithms are trained on labeled data to classify text as positive, negative, or neutral based on patterns and features present in the data.

- Text Preprocessing: The initial step in text analysis, involving data cleaning and transformation to prepare the text for analysis. Techniques include tokenization (breaking text into words), stop-word removal (eliminating common words with little semantic meaning), and stemming (reducing words to their base or root form).

- Google Cloud Platform (GCP): A suite of cloud computing services provided by Google, including virtual machines, storage, and databases. GCP is integrated into this project to store and manage the collected data efficiently.

- Google Cloud Storage Bucket: A storage container in Google Cloud used to store unstructured data, such as text files, images, and videos. In this project, the bucket is utilized to store the collected news articles and other relevant files.

- Google Cloud Datastore: A scalable, NoSQL database provided by Google Cloud, suitable for storing structured data. In this research, it is used to store the collected stock data and sentiment analysis results.

By understanding these crucial terms and concepts, researchers and developers can better grasp the methodologies and techniques employed in the project's implementation, thereby enabling them to effectively conduct sentiment analysis for stock data.

3. Methodology:

The research project follows the following methodology:

- Data Collection:
  - Utilizing the Yahoo Finance API to fetch historical stock data for the specified stock ticker.
  - Web scraping news articles related to the selected stock from reputable news sources.

- Sentiment Analysis:
  - Preprocessing the textual data by removing noise and irrelevant information.
  - Leveraging Natural Language Processing (NLP) techniques to analyze the sentiment expressed in the news articles.
  - Categorizing sentiment as positive, negative, or neutral using machine learning

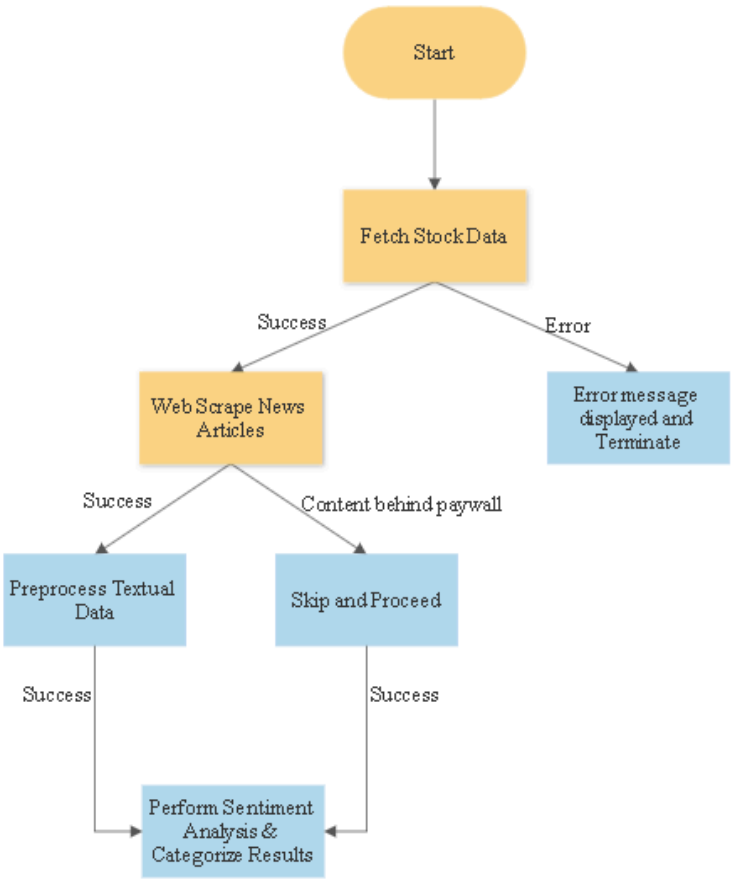
- Data Storage and Integration:
  - Integrating the Python script with Google Cloud for efficient data storage and processing.
  - Storing the collected stock data and analyzed sentiment results in Google Cloud's Storage Bucket and Datastore.

3.1 Data Collection:

- Yahoo Finance API: The Yahoo Finance API is a valuable resource for fetching historical stock data for a given stock ticker. It provides access to key financial parameters, such as the stock's opening price, closing price, and trading volume, allowing for a comprehensive analysis of the stock's performance over time.

- Web Scraping News Articles: Web scraping is employed to acquire news articles related to the selected stock from reputable financial news sources. By extracting textual data from these articles, the sentiment analysis model can analyze and quantify the sentiment expressed in the news, which may influence the stock's market perception and performance.

Here is a flow chart illustrating the steps involved in data collection, including fetching historical stock data using the Yahoo Finance API and web scraping news articles related to the selected stock, would help visualize the process.



3.4 Future Implementations:

While the current research project has achieved significant milestones, there are several potential avenues for future implementations and enhancements:

- Machine Learning Model Refinement: As an extension to the lexicon-based approach, developing and fine-tuning a machine learning-based sentiment analysis model could potentially improve the accuracy of sentiment classification for news articles. Exploring advanced models like Recurrent Neural Networks (RNNs) or Transformer-based models may yield more nuanced sentiment analysis results.

- Sentiment Score Aggregation Strategies: Enhancing the aggregation of sentiment scores from individual words within news articles can be explored. Weighted scoring mechanisms or considering contextual relationships between words may lead to more precise sentiment assessments.

- Real-Time Sentiment Analysis: Implementing real-time sentiment analysis could enable traders and investors to promptly respond to rapidly changing market sentiment. Integrating live news feeds and updating sentiment analysis results in real-time would be a valuable feature.

- Sentiment Correlation with Stock Price Movements: Exploring correlations between sentiment analysis results and actual stock price movements could provide insights into the relationship between market sentiment and stock performance. This analysis may help identify potential patterns and inform trading strategies.

- User Interface Development: Creating a user-friendly web application or graphical user interface (GUI) that allows users to interact with the sentiment analysis tool seamlessly would enhance usability and accessibility.

- Expansion to Multiple Stock Tickers: Extending the project to handle multiple stock tickers concurrently would enable users to analyze and compare market sentiment across various stocks, aiding in diversified investment decision-making.

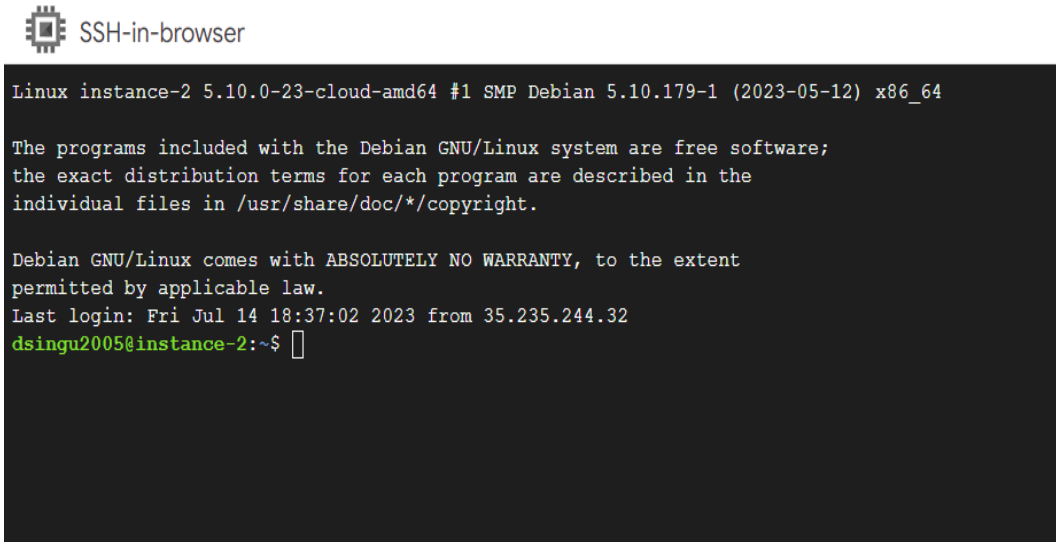
- Sentiment Analysis of Social Media Data: Incorporating sentiment analysis for social media data related to the selected stock ticker could provide additional sentiment insights, as social media platforms often reflect public opinions and market sentiment.

By pursuing these future implementations, the research project can continue to evolve and contribute to the advancement of sentiment analysis for stock data, providing valuable tools and resources for investors and traders in making informed financial decisions.

4. How to Run the Code:**
To execute the Python script and perform sentiment analysis, follow the steps outlined below:
4.1. How to run code (needs to be on both Github and VM) - Part 1

4.1.a. Access the google cloud ‘sentiment analysis project’ and under Computer Engine → VM instances, run instance-2 by clicking on the SSH button in the ‘Connect’ column

4.1.b. After authorizing, your screen will look like this but instead of my user your username will be displayed



4.1.c. Pull the following github repo into the VM, this allows the code to be stored in the VM for running. In case the github repo shows a 404 error, your email may not be added to the repository's shared settings. In order to pull the repo after proper access and authentication, run these commands on the VM in order to get the files:

i. git clone https://github.com/dsingu2005/sentiment-analysis.git

ii. In case the remote authentication fails, follow this tutorial

4.1.d. 'cd' into the correct folder and install the following dependencies with pip like so:

```
pip install yfinance
pip install pandas
pip install bs4
pip install requests
pip install google-cloud
```

4.1.e. Replace the .json' with your own Google Cloud credentials on line 11

```
os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = '.json'
```

4.1.f. Fill in your dedicated ticker at the of the end python file in the executable function:

```
stock = StockInfo("UPDATE TICKER HERE")
```

4.1.g. run the 'stocks&articles.py'

4.1.h. After running the python file this should be the result if the program ran correctly:

```
Success!
stock.csv has been deleted
Success!
news_articles.csv has been deleted
Success!
stock.csv has been created
Success!
news.csv has been created
Success!
news_articles.csv has been created

Success!

news_articles.csv has been created
```

4.2. How to run code (needs to be on both Github and VM) - Part 2

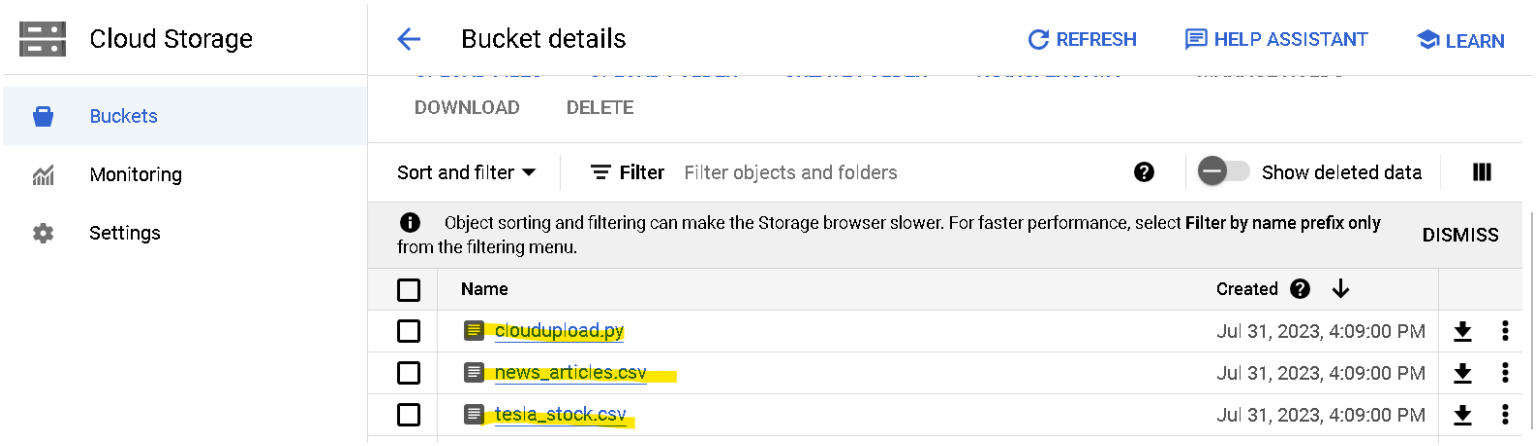
a. After you have completed making any needed changes/optimizations for your specific data needs, the github repo from which you pulled will also contain a 'cloudupload.py' file

b. Navigate to the correct directory and run that file in order to upload your .csv file contents to datastore as well as a copy of your .csv file and .py files to Google Cloud's Storage bucket

c. After running the proper files, a check to make sure that the data and files are properly uploaded to the google cloud's Datastore and Cloud Storage Bucket can be done as follows:

I. Google Cloud Storage:

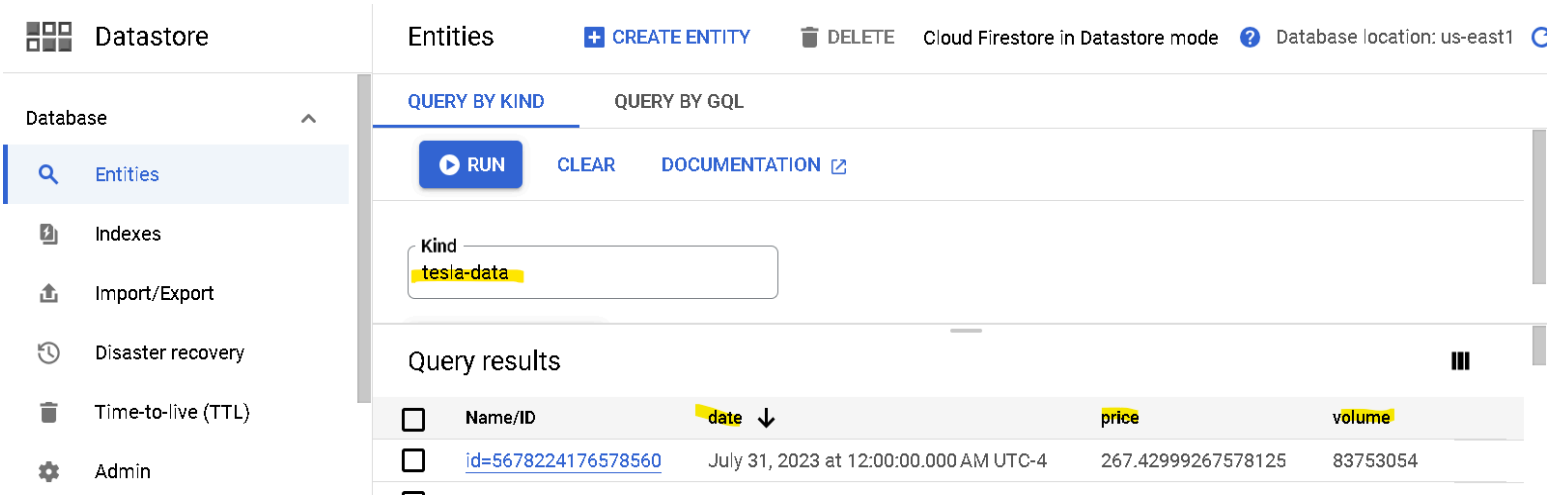
1. Go to the Google Cloud Console
2. In the navigation menu on the left, under "Storage," select "Storage."
3. You should see a list of your storage buckets. Click on the bucket name you used in the script (e.g., 'sentiment-files').
4. You should find the uploaded files listed here. Check if both .csv and 'cloudupload.py' are present.



II. Google Cloud Datastore:

1. Go to the Google Cloud Console:
2. In the navigation menu on the left, under "Datastore," select "Datastore."
3. You will be taken to the Datastore dashboard.

4. In the left-hand menu, click on "Entities." You should see a list of your entities/kinds.
5. If you used the kind 'tesla-data' in the script, find and click on 'tesla-data' otherwise navigate to the kind you used in the script
6. Here, you can check if the data from .csv was added as entities. Look for the properties 'date', 'price', and 'volume' as those are what we pulled from the Yahoo Finance API



Wilton Team

Sentiment Analysis Project - UCONN Stamford & Arjun N Patel

Westhill Team

https://colab.research.google.com/drive/1eKNF5NYOAArdd8kajQv6pASbGZCqDR98?usp=sharing#scrollTo=U7usfYFdchKy

---

Future Design Process for the UConn Team:

1. Data Integration:
- We will integrate the financial data and stock performance metrics from our research with the sentiment analysis results provided by the Wilton Team.
- Ensuring accurate data integration in a suitable format for analysis will be a priority.

3. Feature Engineering:
- We will work with the sentiment analysis results and financial data to identify relevant features for predicting stock price movement effectively.
- Features may include sentiment scores, financial performance metrics in the short term, sector-specific information relative to other companies, and any other relevant data points in order to get more accurate sentiment data.

4. Model Development:
- Building upon the Wilton Team's work, we will develop predictive models using machine learning algorithms and statistical methods.
- We will experiment with different models, considering factors like time series analysis, regression techniques, or ensemble approaches.
- We will focus on enhancing the sentiment analysis model with VADER to improve its accuracy and predictive capabilities. To achieve this, we will implement the following strategies:

1. Increasing Training Data and Learning Samples:
One of the key aspects of improving the sentiment analysis model is to enhance its training data. By providing the model with a larger and more diverse dataset, we can expose it to a wider range of language patterns and financial vocabulary. This will help the model better understand and interpret financial texts, including 10K filings and earnings call transcripts. We will source additional learning samples from various reputable financial sources, academic publications, and other relevant text repositories. By incorporating a vast array of examples, the sentiment analysis model will become more robust and capable of handling various linguistic styles and expressions commonly found in financial documents.

2. Incorporating Historical Data Points:
To obtain a stronger correlation and trendline analysis, we will modify the sentiment analysis model to include historical data points from previous years. Currently, the model primarily analyzes data from recent 10K filings and earnings call transcripts. However, by expanding the temporal scope and considering data from several years prior, the model can identify long-term trends and patterns that may influence stock price movements. This extended historical context will provide a more comprehensive understanding of the market dynamics and allow the model to capture cyclical trends and recurring sentiments that might affect the stock market.

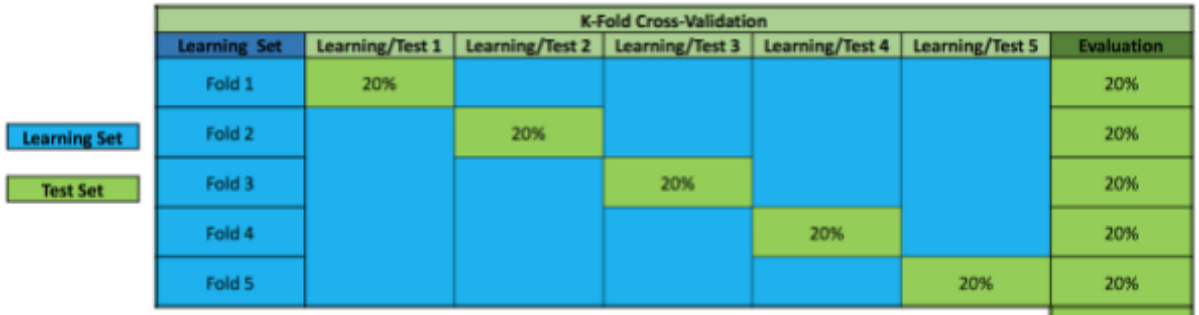
3. Refining the Sentiment Dictionary:
The sentiment analysis model relies on a sentiment dictionary that assigns scores to words based on their emotional polarity. To improve the accuracy of sentiment analysis, we will refine the sentiment dictionary to include a more extensive and domain-specific set of financial terms and expressions. This will involve working closely with financial experts and domain specialists to identify relevant keywords and linguistic nuances that have significant impacts on sentiment analysis within the financial domain. By incorporating a more refined sentiment dictionary, the model can produce more precise sentiment scores, leading to better predictions and insights into stock price movements.

Through these enhancements and refinements, the sentiment analysis model will become a powerful tool for extracting valuable insights from financial texts. The improved model will allow investors and financial analysts to make well-informed decisions based on sentiment trends, historical patterns, and the emotional climate in the financial market. By integrating this advanced sentiment analysis model

with other relevant features and data sources, we will create a comprehensive and robust integrated model for predicting stock price movements with higher accuracy and predictive power in a short term setting. This will pave the way for more informed and data-driven decision-making in the complex world of finance and investment.

4. Model Evaluation:
In the model evaluation phase, we will subject the integrated model to rigorous testing and validation procedures to assess its performance and refine its accuracy. The evaluation process will involve the following key steps:

4.1. Validation Techniques and Cross-Validation:
To ensure the reliability and accuracy of our integrated model, we will implement advanced validation techniques, including k-fold cross-validation, in contrast to the current linear regression testing approach. With k-fold cross-validation, we will partition the dataset into multiple subsets or "folds," and then iteratively train and test the model on different combinations of these folds. This process allows us to obtain a more robust estimation of the model's performance by exposing it to various data distributions and ensuring that it generalizes well across different subsets.



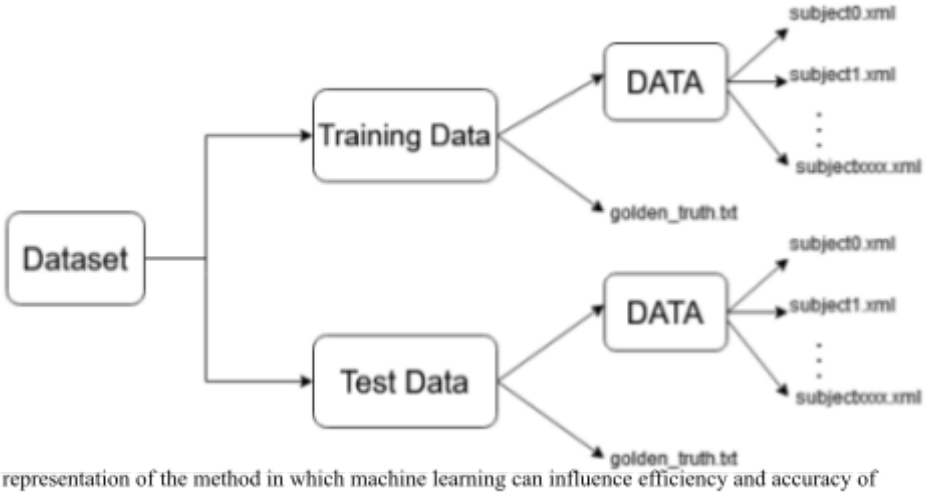
The above image shows the k-fold split in a visual format for better understanding.

By repeatedly shuffling the data and conducting cross-validation, we can effectively mitigate the risk of overfitting, which is a common concern in machine learning models. Overfitting occurs when the model becomes too specific to the training data and fails to perform well on unseen data, leading to false sentiment accuracy. Through cross-validation, we can evaluate the model's ability to make accurate predictions on new and unseen data, providing a more realistic assessment of its predictive capabilities.

By transitioning from linear regression testing to k-fold cross-validation, we aim to overcome the limitations of the current approach and enhance the model's ability to generalize to new market conditions and unseen sentiment data. This will result in a more reliable and accurate sentiment analysis, enabling us to extract valuable insights from financial texts and improve the overall performance of our integrated model. Ultimately, this approach will bolster the model's credibility and applicability in real-world financial scenarios, providing investors and financial analysts with more dependable sentiment-driven predictions for making informed investment decisions.

4.2. Testing on Various Time Periods and Market Conditions:
To validate the model's effectiveness across different market scenarios, we will test it on various time periods, encompassing both bullish and bearish market conditions. This testing approach will allow us to gauge the model's resilience in capturing sentiment trends and predicting stock price movements under diverse economic conditions. We will consider historical data from different years, market cycles, and economic events to ensure that the model performs consistently and accurately across changing market dynamics.

4.3. Addressing Bias and Limitations:
Throughout the evaluation phase, we will also be vigilant in detecting and addressing potential bias in the integrated model. Bias can arise from various sources, such as imbalanced training data or skewed sentiment scores in the sentiment dictionary. We will take necessary measures to mitigate bias and ensure that the model remains fair and unbiased in its predictions. Additionally, we will identify and acknowledge any limitations in the model's capabilities and communicate them transparently in our evaluation report.



A representation of the method in which machine learning can influence efficiency and accuracy of sentiment scores.

By subjecting the integrated model to comprehensive evaluation and continuous refinement, we aim to create a reliable and robust tool for predicting stock price movements based on sentiment analysis and other relevant data points. The insights gained from the evaluation phase will inform our decision-making process in further enhancing the model's performance, ensuring its applicability in real-world financial settings, and supporting informed investment strategies. The final integrated model will empower investors, financial analysts, and decision-makers with valuable sentiment-driven insights to make well-informed and data-driven investment choices in the ever-evolving financial landscape.

5. Fine-Tuning and Optimization, improved trendline analysis, and refined sentiment dictionary:

Fine-tuning and optimization are critical steps in the development of the integrated model for predicting stock price movements based on sentiment analysis. As the UConn Team, we recognize the significance of these steps to enhance the model's performance and ensure its reliability in real-world scenarios. Collaborating with the Wilton Team and other domain experts will be paramount during this phase as their valuable feedback and insights will guide us in refining the model.

Optimizing the model will involve adjusting hyperparameters, such as learning rates, regularization strengths, and the number of hidden layers (if applicable). These hyperparameters significantly impact the model's performance and generalization ability. Through experimentation and cross-validation techniques, we will determine the optimal configuration that yields the best results on multiple subsets of the data.

Feature selection is another critical aspect of optimization. We will conduct a thorough analysis of the features used in the integrated model to ensure that they are relevant and informative. Eliminating irrelevant or redundant features will help reduce noise and improve the model's overall predictive power. By selecting the most influential features, we can enhance the model's ability to capture meaningful patterns in financial text data.

Furthermore, we will implement an improved trendline analysis to provide a more robust understanding of the correlation between sentiment scores and stock price movements. Instead of considering sentiment data from just a few years, we will modify the model to pull data from previous years, extending the historical context of analysis. This longer time frame will enable us to establish a stronger correlation between sentiment and stock prices, yielding a more accurate trendline analysis.

Throughout the fine-tuning and optimization process, we will emphasize the importance of a refined sentiment dictionary. The sentiment analysis is significantly influenced by the words in the dictionary used for scoring. Therefore, we will collaborate with financial experts to implement a more comprehensive and contextually relevant dictionary. This refined dictionary will enable the sentiment analysis to provide more accurate scores, resulting in improved predictions by the integrated model.

6. Results Interpretation:
- Interpretation of the results obtained from the integrated model will be essential to draw meaningful insights into stock price movement prediction.
- Analyzing the relationship between sentiment analysis, financial data, and stock performance will provide valuable findings.