**AI-Driven Stock Prediction with LLMs**

**Technical Specification Document**

**Project Supervisor:** Abhinav Kumar

**Team Members:** Akshat Gosain, Utkarsh Misra, Siddhartha, Aman, Rishabh

**Date:** 10th June 2024

**Project Overview**

The AI-Driven Stock Prediction System is an innovative project that aims to revolutionize stock market predictions. Leveraging the power of Large Language Models (LLMs), the system summarizes the latest market news and performs sentiment analysis to predict stock price trends. This project is expected to provide valuable insights to investors, enabling them to make informed investment decisions.

The system takes the name of a stock as input, identifies the relevant market domain, and extracts the latest news using beautiful soup and Google search results. The news is then summarized using a fine-tuned LLM. The system performs sentiment analysis on the summaries to extract features, which are stored and used by a binary classifier to predict whether the stock price will go up or down the next day.

This project is scheduled for completion in 1.5 months. The team is committed to delivering a high-quality, reliable, and efficient system that meets the needs of its users.

**Table of Contents**

1. **Introduction** - Page 3
   * Project Overview
   * Purpose of the Document
   * Scope of the Project
2. **System Overview** - Page 4
   * System Architecture
   * System Components
3. **Stock Input and Domain Identification** - Page 5
   * Input Interface
   * Domain Identification Process
4. **News Extraction** - Page 6
   * Beautiful Soup Extraction
   * Google News Extraction
5. **News Summarization using LLMs** - Page 7
   * LLM Model Description
   * Summarization Process
6. **Sentiment Analysis** - Page 8
   * Sentiment Analysis Model
   * Binary Classification Process
7. **Output and Reporting** - Page 9
   * Output Interface
   * Report Generation
8. **System Requirements** - Page 10
   * Hardware Requirements
   * Software Requirements
9. **Implementation Timeline** - Page 11
   * Project Milestones
   * Estimated Completion Dates
10. **Conclusion** - Page 12
    * Summary of the Project
    * Future Work

**Project Overview**

**Brief Description of the Project**

This project aims to develop a **Stock Market Prediction System** that leverages **Language Model-based Learning Machines (LLMs)** for summarizing the latest market news and performing sentiment analysis. The system will take a stock name as input, identify the relevant domain, extract news using Beautiful Soup and Google, summarize it using an LLM, and use binary classification for sentiment analysis. The goal is to provide users with accurate and concise summaries of the latest news related to their chosen stocks, along with a sentiment analysis that could potentially guide their investment decisions.

**Objectives of the Project**

1. **Data Collection**: Gather and organize relevant stock market news from various sources such as Google.
2. **News Summarization**: Implement an LLM to summarize the collected news articles, providing users with concise and understandable summaries.
3. **Sentiment Analysis**: Apply binary classification techniques to perform sentiment analysis on the summarized news, categorizing them as either positive or negative.
4. **User Interface**: Develop a user-friendly interface where users can input the name of a stock and receive the summarized news and sentiment analysis.
5. **Performance Evaluation**: Evaluate the accuracy of the news summarization and sentiment analysis, and the overall effectiveness of the system in aiding users’ investment decisions.

**Expected Duration of the Project**

The project is expected to be completed in **1.5 months**. The timeline is as follows:

* **Weeks 1-2**: Data collection and organization
* **Weeks 3-4**: Implementation of news summarization using LLMs
* **Weeks 5-6**: Implementation of sentiment analysis and development of the user interface
* **Week 7**: Performance evaluation and final adjustments

**System Architecture**

The system architecture for the stock prediction project using LLMs for summarization of latest market news and sentiment analysis is designed to be modular and scalable.

**1. Input Interface:**

The system starts with an **Input Interface** where the user can input the name of the stock they are interested in. This interface is designed to be user-friendly and intuitive, allowing users to easily input the required information.

**2. Domain Identification:**

* Once the stock name is input, the system proceeds to the **Domain Identification** module. This module identifies the domain or industry of the stock, which is crucial for the subsequent news extraction process. The domain identification is done using a pre-trained machine learning model.

**3. News Extraction:**

* The **News Extraction** module is responsible for extracting the latest news related to the stock from various sources such as Google. The news extraction is done using web scraping techniques like beautiful soup.

**4. News Summarization:**

* The extracted news articles are then passed to the **News Summarization** module. This module uses a Language Model (LLM) to summarize the news articles into concise summaries. The LLM is trained on a large corpus of financial news articles to ensure the summaries are accurate and relevant.

**5. Sentiment Analysis:**

* The summarized news articles are then passed to the **Sentiment Analysis** module. This module uses a binary classification model to determine the sentiment (positive or negative) of the news. The sentiment analysis is crucial for the final stock prediction.

**6. Stock Prediction:**

* Finally, the **Stock Prediction** module uses the sentiment analysis results along with other factors such as historical stock data and market trends to predict the future performance of the stock.
* Here is a flowchart illustrating the system architecture:
* User Input -> Domain Identification -> News Extraction -> Summarization -> Sentiment Analysis -> Stock Prediction
* Each arrow represents the flow of data from one module to the next. Each module is designed to be independent and interchangeable, allowing for easy upgrades and modifications.

**Technical Specifications**

## 1. Input

The system takes the **stock name** as input from the user. The stock name is used to identify the domain and extract relevant news articles.

## 2. Process Flow

The process flow of the project can be broken down into the following steps:

### **Domain Identification:**

Once the stock name is provided, the system identifies the domain of the stock. This is crucial for extracting relevant news articles.

### **News Extraction:**

The system extracts news articles related to the identified domain from **Google**. The news articles provide the latest market news about the stock.

### **News Summarization:**

The extracted news articles are then summarized using a **Language Model (LLM)**. The LLM generates a concise summary of the latest market news.

### **Sentiment Analysis:**

The system performs sentiment analysis on the summarized news using a **binary classification model**. The sentiment analysis provides an understanding of the market sentiment towards the stock.

## 3. Output

The output of the system is a **summary of the latest market news** and the **sentiment analysis result** for the given stock.

## 4. Technologies and Tools Used

The project uses various technologies and tools for different parts of the process:

### **News Extraction:**

For news extraction, the system uses **web scraping tools** to extract news articles from Google.

### **News Summarization:**

The news articles are summarized using a **Language Model (LLM)**. The LLM is trained on a large corpus of text and can generate a concise summary of the news articles.

### **Sentiment Analysis:**

The sentiment analysis is performed using a **binary classification model**. The model is trained on a dataset of news articles and their associated sentiments.

# Model Training and Evaluation

## 1. Model Training

The model training process involves two main steps: **Summarization of Market News** and **Sentiment Analysis**.

### **1.1 Summarization of Market News**

For the summarization of market news, we will use a pre-trained **Language Model (LM)**. The LM will be fine-tuned on a corpus of financial news articles. The input to the model will be the raw text of the news article and the output will be a concise summary of the article.

The model will be trained using a variant of the **transformer architecture**, which has shown state-of-the-art performance on a variety of NLP tasks. The training process will involve minimizing a loss function that encourages the model to generate summaries that are both fluent and informative.

### **1.2 Sentiment Analysis**

For sentiment analysis, we will use a **Binary Classification model**. The input to this model will be the summarized text from the previous step and the output will be a sentiment score, indicating whether the sentiment of the text is positive or negative.

The model will be trained on a labelled dataset of financial news summaries, where each summary is associated with a sentiment score. The training process will involve minimizing a loss function that encourages the model to correctly predict the sentiment of the text.

Here we will also add other features under emotional spectrum- hate , anger, happy, sad etc. these feature will better help us in knowing the sentiment of the data. This will help us understand the deeper insights of the data and will improve the accuracy and potential of our model.

### **1.3 Machine Learning Algorithms**

We will be training our model on different machine learning algorithms like- KNN(k-nearesr neighbour ), SVM(Support vector Machine), Classification, Regression, Random forest, XGBoost and many more and after that we will compare all the algorithm and will take the one which will have the best accuracy.

## 2. Model Evaluation

### **2.1 Evaluation of Summarization Model**

The performance of the summarization model will be evaluated using the **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric**, which compares the generated summaries with reference summaries to compute scores for precision, recall, and F1-score.

### **2.2 Evaluation of Sentiment Analysis Model**

The performance of the sentiment analysis model will be evaluated using standard classification metrics such as **accuracy, precision, recall, and F1-score**. In addition, we will also compute the **Area Under the Receiver Operating Characteristic (AUROC)** to measure the model’s ability to distinguish between positive and negative sentiment.

**Performance Metrics**

1. **Accuracy**: This is the most straightforward metric. It is simply the proportion of correct predictions made out of all predictions. Accuracy is a good measure when the target variable classes in the data are nearly balanced.
2. **Precision**: Precision is the ratio of correctly predicted positive observations to the total predicted positives. High precision relates to the low false positive rate. It is a good measure to determine when the costs of false positives are high.
3. **Recall (Sensitivity)**: Recall is the ratio of correctly predicted positive observations to all actual positives. It is a good measure to determine when the costs of false negatives are high.
4. **F1-Score**: The F1 Score is the weighted average of Precision and Recall. It tries to find the balance between precision and recall. It is a good measure when you have an uneven class distribution.
5. **Area Under the ROC curve (AUC-ROC)**: ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.
6. **Mean Squared Error (MSE)**: MSE is the average of the squared differences between the predicted and actual values. It is a popular metric for regression problems.
7. **Root Mean Squared Error (RMSE)**: RMSE is the square root of the average of the squared differences between the predicted and actual values. It is more sensitive to larger errors than MSE.
8. **Mean Absolute Percentage Error (MAPE)**: MAPE measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error.
9. **Sentiment Accuracy**: This is specific to the sentiment analysis part of your project. It measures the proportion of correct sentiment predictions made out of all sentiment predictions.
10. **News Summarization Quality**: This is a subjective metric that can be evaluated through user feedback. It measures the quality, relevance, and comprehensibility of the news summaries generated by the LLM.

**Future Work**

**1. Enhancements**

* **Sentiment Analysis Improvements**: We plan to enhance the sentiment analysis model by incorporating more sophisticated Natural Language Processing (NLP) techniques. This could include the use of transformer-based models like BERT or GPT-3 for better understanding of the context in news articles and headlines.
* **Incorporating More Data Sources**: Currently, we are extracting news from RSS feeds and scraping Google for headlines. In the future, we aim to incorporate more diverse data sources like financial reports, social media chatter, and expert analysis reports.
* **Real-time Analysis**: We aim to develop a real-time analysis feature that can provide instant sentiment scores as soon as a news article is published. This will allow for more timely stock predictions.
* **Anomalies Detection**: A sudden spike or drop in stock price can occur independent of news sentiment. Factors like unexpected earnings reports, mergers and acquisitions, or industry-wide events can significantly impact prices.
* **Fake News :** We will plan to build a feature which will help us find the fake news while collecting the news from different sources. This will help us increase our accuracy.

**2. Scalability**

* **Handling Larger Data Volumes**: As we incorporate more data sources, the volume of data will increase significantly. We plan to use more efficient data processing and storage solutions to handle this increase.
* **Automating the Fine-Tuning Process**: The process of fine-tuning the LLM with historical data and latest news is currently semi-automatic. We aim to fully automate this process to save time and resources.
* **Expanding to More Stocks**: Currently, the project is focused on a limited number of stocks. In the future, we aim to expand the scope to include a larger number of stocks from various sectors and markets.

These enhancements and scalability improvements will significantly improve the performance and utility of our stock prediction project. They will allow us to provide more accurate predictions, handle larger volumes of data, and cover a wider range of stocks. This will ultimately provide more valuable insights for our users and help them make more informed investment decisions.

Please note that these enhancements and improvements will require additional resources and time to implement. We will need to plan accordingly to ensure that we have the necessary resources and expertise to successfully implement these changes.

**Conclusion**

**Summary of the Project**

The project aimed to predict stock prices using Language Learning Models (LLMs) for summarizing the latest market news and sentiment analysis. The approach involved extracting news by scraping Google for articles, and fine-tuning an LLM with historical data and the latest news. The project also incorporated binary classification for sentiment analysis to predict stock price movements.

**Potential Impact**

The potential impact of this project is significant:

1. **Improved Accuracy**: By incorporating the latest news and sentiment analysis into the prediction model, the project has the potential to significantly improve the accuracy of stock price predictions.
2. **Real-Time Analysis**: The use Google scraping allows for real-time analysis of market news, which can be crucial in the fast-paced world of stock trading.
3. **Informed Decision Making**: The summarization of market news can help traders and investors make more informed decisions, potentially leading to increased profits.
4. **Scalability**: The project’s methodology can be applied to other financial markets, making it highly scalable.
5. **Innovation**: The project represents an innovative approach to stock price prediction, combining several different techniques (LLMs, sentiment analysis, web scraping) in a novel way.

In conclusion, this project has demonstrated the potential of combining LLMs, sentiment analysis, and real-time news updates to predict stock price movements. The success of this project could pave the way for more sophisticated and accurate stock prediction tools in the future. It’s an exciting step forward in the field of financial technology.

**References**

1. Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2, no. 1 (2011): 1-8.

2. Nguyen, Thien Hai, and Kiyoaki Shirai. "Topic modeling based sentiment analysis on social media for stock market prediction." In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 135-140. 2015.

3. Patel, Jigar, Sahil Shah, Priyank Thakkar, and K. Kotecha. "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques." Expert Systems with Applications 42, no. 1 (2015): 259-268.

4. Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." ACM Transactions on Information Systems (TOIS) 27, no. 2 (2009): 1-19.

5. Devitt, Alan, and Khurshid Ahmad. "Sentiment polarity identification in financial news: A cohesion-based approach." In Proceedings of the 45th annual meeting of the association of computational linguistics, pp. 984-991. 2007.

6. "RSS Feeds." RSS 2.0 Specification. [Online]. Available: https://www.rssboard.org/rss-specification

7. "Beautiful Soup Documentation." Beautiful Soup 4.9.0 documentation. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

8. "Google News." Google News. [Online]. Available: https://news.google.com/

9. "PyTorch Documentation." PyTorch. [Online]. Available: https://pytorch.org/docs/stable/index.html

10. "Scikit-learn: Machine Learning in Python." Scikit-learn 0.24.1 documentation. [Online]. Available: https://scikit-learn.org/stable/

**Appendices**

**Appendix A: RSS Feed Sources**

This appendix contains a list of RSS feeds if being used to extract the latest market news. The feeds are from various financial news websites and blogs that provide up-to-date information about the stock market.

**Appendix B: Google Scraping Methodology**

This appendix details the methodology used for scraping Google for headlines related to the stock market. It includes the search queries used, the frequency of scraping, and the method of extracting and storing the headlines.

**Appendix C: LLM Fine-Tuning Details**

This appendix provides information about the fine-tuning process of the Language Model. It includes details about the historical data used, the preprocessing steps, the training parameters, and the evaluation metrics.

**Appendix D: Sentiment Analysis Model**

This appendix describes the binary classification model used for sentiment analysis. It includes information about the model architecture, training data, performance metrics, and usage in predicting stock price movements.

**Appendix E: Project Timeline**

This appendix presents a detailed timeline of the project, including the start and end dates of each phase, major milestones, and the expected completion date of the project.

**Appendix F: Team Members and Responsibilities**

This appendix lists all the team members involved in the project and their respective responsibilities. It also includes their contact information for easy reference.