**Financial Text Analysis: Using NLP to Forecast Stock Prices**

**Project Domain & Goals**

This study aims to utilize Natural Language Processing (NLP) to address the drawback of technical analysis in stock prediction, emphasizing historical stock price. Our team suggests using NLP to integrate technical analysis with fundamental analysis. Fundamental analysis typically requires a thorough manual examination of financial documents, including annual reports (10-K), quarterly reports (10-Q), and current reports (8-K). Our strategy utilizes NLP to extract relevant features from these financial records and integrate it with technical analysis models like the moving average (MA) and long short-term memory (LSTM). We'll assess whether incorporating these essential features can make the initial models perform better. Due to the time constraint, we will focus on 10 largest stocks in Dow Jones with 5-year historical data. Because the stock market is sensitive to information in these financial documents, we believe that by integrating fundamental and technical analysis, we may improve stock market forecast performance.

**Related Works**

Several machine learning models have been used to analyze historical financial data and predict the stock price to support short-term trading strategies. According to the efficient market hypothesis, the current stock price reflects all information in the market, and if new information becomes available, the price will respond accordingly. This means that not only numerical financial data but other forms of information such as news, articles, and financial documents can affect the stock price as well. Research suggests that news and articles correlate to the behavior of stock price (Gidófalvi, 2001), and the sentiment from the

news is also used to help investors trending decisions (Mishev et al., 2020).

Previously, machine learning models were used to process historical financial data and textual data separately, and the interconnection between numerical and textual information that could affect the stock price is overlooked. Recently, the joint effect of various types of information has been studied. In Kalra (2019), to predict the Indian stock market, news articles and historical data were used. News articles were preprocessed using tokenization, transformation, stop word removal, Term Frequency, and Inverse Document Frequency (TF-IDF). Sentiment analysis has been performed on the preprocessed dataset by the naive Bayes Classifier. Then, sentiment values and historical data were integrated by date to form predictors and those predictors were given to the Machine Learning models such as KNN, Neural network, Super vector Machine, and Naive Bayes for stock prediction.

In Guo (2020), news articles and price history were used to predict stock prices. Sentiment analysis has been done on the textual data (news) using two python libraries, VADER (Valence Aware Dictionary and Sentiment Reasoner) and TextBlob, which changes text into sentiment scores. After that, all the essential features from news and historical stock basic features have been used in the LSTM neural network to predict close prices and stock returns.

For this project, financial documents will be used instead of news articles, and multiple features extractions techniques, such as Word2Vec, N-grams, FINBERT, and topic modeling will be used and compared to see which will work best. Moreover, we will use machine learning models such as LSTM and MA to predict the stock price to see if including information from financial documents will improve the model's performance.

**Datasets**

In this project we will use three widely available datasets:

● Financial documents such as annual reports (10-K), quarterly reports (10-Q), and current reports (8-K) can be obtained from the Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system

(https://www.sec.gov/edgar/searchedgar/com panysearch).

● Federal Reserve Economic Data, FRED (https://fred.stlouisfed.org/fred-mobile/index. php), an online database consists of economic time series data from scores of national, international, public, and private sources.

● Yahoo Finance (https://finance.yahoo.com/) provides free access to real-time stock quotes for most US stocks, as well as historical data for numbers of companies.

For each dataset, the first step will be to parse the documents and perform data cleaning and preprocessing through standard NLP methods to remove noise and convert it into a standardized format. This includes tokenization, stop word removal, stemming, or lemmatization. Once the pre-processing is complete, the next step will be to extract the relevant features from the text data. This can be done using techniques such as TF-IDF, Word2Vec, or Sentiment Analysis.

In case the desired results are not achieved, the alternate approach will be to obtain a dataset that includes news, and articles, and scrape these websites to collect data related to the companies of interest. Potential sources will be Financial news websites such as Bloomberg, Reuters, and CNBC; Social Media Platforms like Twitter, where people share information about stocks and companies.

**Technical Challenge**

The challenges on the processes of data acquisition, data preprocessing, and textual features extraction are addressed in the first

phase of the project while improving model performance by experimenting different algorithms and techniques as well as understanding the impacts of adding textual financial data into the model will be analyzed in the second phase.

For the data acquisition part, we planned to extract features from financial reports of the selected public companies, hence we need to develop a code to crawl the website to download hundreds of these documents, and process them in machine-translatable formats like HTML or PDF. During the preprocessing step, a logic to aggregate different datasets as well as define our target variable will be discussed. Subsequently, we need to extract features from the financial reports using different NLP-related techniques e.g. TF-IDF, Word2Vec, or NLTK’s VADER module to extract sentiments of the texts.

Given that the stock market prediction is a difficult task in terms of getting high accuracy due to various factors and uncertainties, the main expectation of our project is to develop a model which incorporates both fundamental and technical features to give better prediction than the traditional baseline model which normally considers only technical features. Even though it is not guaranteed to boost the performance up, we will analyze whether different parts or techniques implemented led to positive or negative effects on the prediction. With this analysis, further studies could be conducted on these areas that have the potential and lead to improved stock analysis in the future.

**Work division**

Paniti will be responsible for data acquisition. Vorapoom will be responsible for Data processing. Thanawan, Rishitha, and Neeharika will be responsible for Model training and evaluation.

**References**

1. Győző Gidófalvi. 2001. Using News Articles to Predict stock Price Movements.

Semantic Scholar. https://people.kth.se/~gyozo/docs/financial prediction.pdf

2. Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. 2020. Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. IEEE. https://ieeexplore.ieee.org/abstract/document /9142175/authors#authors

3. Yuqiao Guo. 2020. Stock Price Prediction Based on LSTM Neural Network: the Effectiveness of News Sentiment Analysis. IEEE.

https://ieeexplore.ieee.org/document/943481 0

4. Sneh Kalra, Jay Shankar Prasad. 2019. Efficacy of News Sentiment for Stock Market Prediction. IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?t p=&arnumber=8862265