Using NLP and Supervised ML to Predict Changes in Stock Price from Company Earnings Call Transcripts

Emily Wang, Xinyu Wu, Kelly Phalen, Marco Tortolani, Qi Li

https://github.com/Qi-Li3/DS3500-Final-Project

I. Motivation

People make choices on what to buy or avoid based on their opinions about companies on a daily basis, which influences a company's earnings. The same principle applies to market participants since they form opinions about the companies they are invested in based on reported news and publications of the companies' earnings. The participants form market sentiment and make decisions leading to buying and selling stocks. This is significant because the psychology of market participants has really only been studied in the last two decades. Now, however, it is even easier to access news and data like earnings reports. We want to determine how much market sentiment of investors affects stock prices. Our model will predict when stocks will rise or fall based on companies' earnings calls, which then shows how people reacted. This could allow investors to determine if a certain investment will be profitable and companies to foresee their future gains or losses. This idea is worthy of a whole semester project because it requires acquiring data from multiple sources, creating a database, using natural language processing on our data, and using machine learning to create our stock prediction model. Our project will require extensive research and advanced programming techniques to execute. Future recruiters will be interested to see the variety of skills we used on our model that determines stock prices based on publications. Additionally, working and contributing to a project with a team through GitHub is important to recruiters, as it is how many companies actually operate. The project highlights our interdisciplinary thinking. Furthermore, our project is interesting because it relies heavily on the psychology of market participants, which we cannot directly know, but rather we are obtaining from sentiment analysis of text data. Overall, we hope to understand how publications like earnings reports affect stocks and build a successful model to predict future stocks.

II. Goals and Objectives

In this project, we hope to accurately analyze company earnings calls in order to predict the rise and fall of stock prices. We hypothesize that positive sentiment about a company will correlate with a rising stock price while negative sentiment will correlate with falling stock prices. First, we will create reusable methods that utilize APIs to obtain quarterly earnings call transcripts and stock prices. The earnings call data will be cleaned and added to a PostGreSQL database. Using this data, we will create a NLP and supervised learning pipeline to process and make predictions. Potential techniques include Bag-of-Words, TF-IDF vectorizer, and n-gram. For the supervised ML model, we will test out algorithms including Multinomial Naive Bayes, Support Vector Machine, and Linear Regression. Stock price data will be used to train the model, which will produce a binomial output of 'increase' or 'decrease'. We plan to conduct hyperparameter tuning and model-based feature selection as well as GridSearch with cross validation to gauge model performance and further optimize our ML models. Our pickled models can be useful for future developers or investors looking to invest in the market. External users can easily access the database and add to it for future projects. We also plan to develop reusable API access tools for future developers to get a company's financial data by simply entering a

ticker symbol. While we will only look at the top 10-20 companies listed on the NASDAQ (by market capitalization), future developers can extend this to other companies.

III. Data Sources

Since our project compares financial reports to stock performance, our job in sourcing data will be to acquire company earning transcripts and company stock history. Companies such as Finnhub (https://finnhub.io/) offer APIs in order to acquire earnings call transcripts. Acquiring stock data is made easy through the Yahoo Finance API (https://yfapi.net). This data will not be stored in the database as it will not be regularly used to train the NLP model.

IV. Platform Architecture

Our final project will produce three primary architectural components: a reusable set of tools to access and clean API text data, a PostGreSQL database to store earnings call transcripts, and a pipeline of NLP tools and supervised machine learning models that can predict whether a stock price will go up or down. The library of reusable tools will include API connection functions, text cleaning methods, and will add the data to the database. The PostGreSQL database will be used for storing company's quarterly earnings call transcripts that are cleaned and ready to input into the models. The trained models will use a variety of NLP techniques and supervised learning models to produce an output (as described in the 'Goals and Objectives' section). The pipeline of models will be pickled for easy future use.