CSCI-4502+5502

Peter Ly
Peter.Ly@colorado.edu
CSCI-5502

Vijay Kumar Poloju
Vijay.Poloju@colorado.edu
CSCI-5502

Harrison Fagan
hafa@9887@colorado.edu
CSCI-4502

Andrew Connell
anco2818@colorado.edu
CSCI-4502

Abstract

The report accompanying the initial project proposal presentation and announcement. This document elaborates primarily on the related works to the problems we want to solve and provides more detail on our plan for the project.

ACM Reference Format:
Peter Ly, Vijay Kumar Poloju, Harrison Fagan, and Andrew Connell. 2018. CSCI-4502+5502. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Gathering insight from financial data has been a key area for many researchers for a number of years since the advent of modern statistical and computational methods. From as early as the 1900s researcher have been trying to draw insights from the collected market data. There have been a multitude of different approaches, but with modern statistical, and computational methods we have seen an explosion in this field of research. With new natural language processing (NLP) Techniques it becomes possible to use a new dense source of data within these analyses, text. Automated processing of text can give us new indicators that, without, could have been overlooked or too nuanced to be found manually. Along with new NLP techniques the ever increasing amount of computing power paired with newer statistical methods provides an opportunity to revisit analyzing correlations between stock prices. Knowing these trends and correlation could add a whole new array of features to tap in the future for trying to solve problems like financial forecasting and market analysis. Being able to accurately and quickly assess the trend of assets from text would be invaluable. Our work will focus on using these techniques together to understand whether they offer a new path for financial modeling.

2 Related Work

There is a wide body of literature concerning aspects of company performance prediction. In this section, we identify studies which address problems similar to the problems we want to answer, what techniques were used in the particular study, and how our work differs from their study. We generally categorize the related work by the general purpose they have.

2.1 Trend Analysis in Bsinesses

In 2021, Saritas, Bakhtin, Kuzminov, and Khabirova [7] studied current trends in big data and mobile commerce, highlighting its contemporary relevance. The researchers aimed to analyze how big data technologies can augment business trends, particularly in mobile commerce. They explored how businesses can leverage mobile platforms to enhance customer experiences and increase engagement through big data analytics primarily by using several technologies: dimensionality reduction, feature engineering, text mining and real-time data stream processing. Their work focuses on using big data analytics to augment business trends in mobile commerce and thereby obtain actionable insights to enhance mobile commerce strategies and customer engagement whereas our project focuses on NLP and sentiment analysis applied to company earnings reports to analyze how company outlooks influence stock price movements and investigate unusual correlations between companies.

In June 2021, Pantenburg and Nordbring studied possible correlation between the tone of CEO letters and the subsequent performance of their companies [5]. In order to find the tone of the CEO letters, they analyzed a range of CEO letters using a collection of positive and negative words. They gathered stock data from the Nikkei 225, S&P 500, and OMXSPI indexes. To identify correlations, they applied Spearman correlation analysis. In contrast, our work does not focus on the tone of CEO letters; instead, it will concentrate on the revenue and general technical direction of the companies.

In March 2023, Rajan, Ramella, and Zingales conducted research aimed at identifying trends in a company's stated goals and analyzing whether these companies made efforts to fulfill those goals [6]. They used natural language processing to extract various goals and examined how these goals changed over time. In contrast, our work seeks to integrate stock data alongside various company's goals to make predictions, while their study focused solely on the goals and their changes over time.

2.2 Stock Performance Prediction

In 2015, Wafi, Hassan, and Mabrouk in a comprehensive survey of the field of fundamental analysis [8] define the concept of fundamental analysis as "Knowledge of the rules and fixed steps access to its objectives of determining the intrinsic value of shares in stock markets, through a general framework to study the expected economic forecasts, leading to sectors which generate an increase in

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

sales and profits, therefore measure strength financial companies, efficiency of management and business opportunities based on historical financial statements and current conditions. Thus determine the stock fair value, and then compare them to market values resulting from interactions of supply and demand, to identify investment opportunities (profit or loss)." Additionally they describe a few models used within fundamental analysis including dividend discount, multiple dependent models, discounted cash flow models, residual income valuation model.

In 2010, Buda [2] describes the deficiencies with the usual statistical method for calculating coefficients between samples is insufficient for stocks, as it provides an unconditioned estimate. Along with this finding the correlation raises another issue which is how long the window for trying to find this correlation should be. Buda describes methodology for creating more accurate estimates for the correlation coefficient as well as the most effective time window for calculating correlations between stocks. His research concluded that this window was around 3 to 4 months.

In 2016, Kim [4] performed time-lagged correlation analysis between stock prices of two companies which were known to be in the same sector over a period of 47 days in order to determine if predicting stock prices using correlation coefficients would be effective in obtaining profit via purchasing and selling stocks. The work performed by Kim was primarily statistical, did not consider using data from shareholder letters, and was short-term whereas our study focuses on combining data from more sources, larger time periods, and correlations outside of stock prices which can be derived.

In 2018, Guo, Zhang, and Tian [3] performed a correlational analysis between stock prices of several companies in China, focusing on non-linear measures of correlation (namely, the mutual information measure). They use this measure to identify positive and negative correlations between different stocks and develop a stock network encoding the relationship between different stocks in China. This study uses only stock data whereas our study will focus on using other sources of data. Furthermore, while their study did classify the stocks by industry, it did not determine whether there were other classifications for the stocks which would perform better than classification by industry, whereas our study will focus on better understanding stocks using a more nuanced classification than industry.

In 2024, Balasubramanian, P, Badarudeen, and Sriraman conducted a survey on the technologies used for individual stock price predictions in the time period 2000 to 2020 [1]. Generally, there were statistical and machine learning techniques used for this purpose. These include natural language processing (NLP), various machine learning models (for example, support vector machines, random forests, and neural networks), time-series analysis, and long short-term memory (LSTM). Many of these studies focused only on individual prediction for companies, with little consideration for cross-company correlations.

3 Proposed Work

Our work focuses on solving several key problems: understanding the impact of business decisions on stock performance, identifying patterns and trends across multiple companies, investigating unusual correlation between companies, and predicting future stock performance. Our proposed work includes sentiment analysis of earnings reports, time series analysis of stock prices, identifying correlations between companies and building predictive models for stock movement.

We have four overarching (related) goals for this project. In each subsection, we identify an overarching goal for the project, and propose the necessary data for each goal to be completed.

3.1 Measurement of Impact of Business Decisions on Stock Performance

We will analyze earnings reports to determine how specific business decisions influence stock prices, using both statistical and sentiment analysis. To conduct this analysis, we will need both historical earnings reports and stock price data.

3.2 Identification of Patterns and Trends across Multiple Companies

By applying natural language processing and time-series analysis, we will extract patterns from earnings reports and stock trends, allowing us to identify broader trends across companies. To conduct this analysis, we will need both historical earnings reports and stock price data.

3.3 Potential Partnership Identification

Through correlation analysis, we aim to detect unusual relationships between companies that could suggest hidden partnerships or collaborations. To conduct this analysis, we will need historical stock price data and publicly available press announcements.

3.4 Future Stock Performance Prediction

Using predictive modeling techniques such as long short-term memory, we will forecast future stock prices based on historical data and current business decisions. To conduct this analysis, we will need both historical earnings reports and stock price data.

4 Evaluation

In this section, we detail how we will validate that our methods work and how we will determine the accuracy of our methods mentioned in our proposed work.

4.1 Measurement of Impact of Business Decisions on Stock Performance

To validate our method, we will select a small collection of related businesses with similar significant directions and in particular market sector, and verify that the significant direction is a better indicator of stock performance than the market sector. In particular, we ideally should see that the emphasis on the significant direction in shareholder letters is positively correlated to stock performance in this group.

To measure the accuracy of our method, we will select a collection of companies, and use historical shareholder letters to identify the significant directions of the companies. Then, we will compare the emphasis on the direction stock performance to stock performance for each company.

4.2 Identification of Patterns and Trends across Multiple Companies

To validate our method, we will select a small number of companies which have made similar directions in industry recently, have maintained similar stock performance, but have not necessarily had similar history in the past. For these companies, we will analyze the content of their shareholder letters to verify that they are progressing in similar directions.

To measure the accuracy of our method, we compare the patterns and trends we collect over several companies to the patterns and trends proposed by business consulting firms (for example, McKinsey). Generally, we seek to find patterns and trends which are not mentioned by consulting firms and span across multiple industries.

4.3 Potential Partnership Identification

To validate our method, we will select a subset of companies with known partnerships and known non-partnerships. We expect that the correlation between stock prices and shareholder letters will be higher for companies with partnerships than for companies without partnerships.

To evaluate the accuracy of our method, we will select a large set of companies and examine their recent performance to determine their correlation. Then, for pairs of companies which exhibit high correlation, we will verify whether a known partnership exists between the companies.

4.4 Future Stock Performance Prediction

To validate our method, we will use past stock data of a few companies to generate our prediction, and then verify that the extrapolation from the past data aligns with later past data.

To evaluate the accuracy of our method, we will use a large set of companies to which to verify that extrapolation from past data aligns with later observed stock behavior.

5 Milestones

In this section, we propose a tentative list of milestones that must be completed in order to complete the project. We also include estimations on the time to complete each milestone, though these may need to change based on data format and API access.

5.1 Data collection

This task should be done in two phases. One phase will be to gather stock data, which should be pretty easy. This should be completed in the first week. The second phase involves collecting information from shareholder letters, which will be more challenging. We will use natural language processing to extract data from these letters. A basic algorithm should be finished within the first few weeks, with changes made later.

5.2 Data cleaning

After we collect the data, we should change it into a usable format. This task may require us to go back to the collection process to gather additional data or refine our searches. This task should take about one to two weeks.

5.3 Data integration

Following data cleaning, we will integrate our datasets by grouping them according to company. This task should take around one to two weeks.

5.4 Case study validation

This task relies only on a small subset of our data. Here, we will develop one or more models to analyze our data, which will then be tested on a small sample. We estimate that this process will take four to five weeks.

5.5 Complete modelling

In this phase, we will expand our initial model or models to use the entire dataset. This task depends on the completion of all prior tasks, and we think it may take one to two weeks.

References

[1] Prakash Balasubramanian, Chinthan P, Saleena Badarudeen, and Harini Sriraman. 2024. A systematic literature survey on recent trends in stock market prediction. *PeerJ Comput. Sci.* 10 (Jan. 2024), e1700.

[2] Andrzej Buda. 2011. Life time of correlation between stocks prices on established and emerging markets. arXiv:1105.6272 [q-fin.GN] https://arxiv.org/abs/1105.6272

[3] Xue Guo, Hu Zhang, and Tianhai Tian. 2018. Development of stock correlation networks using mutual information and financial big data. *PLOS ONE* 13, 4 (04 2018), 1–16. https://doi.org/10.1371/journal.pone.0195941

[4] Sungil Kim. 2016. A Cross-Correlation-Based Stock Forecasting Model. https://api.semanticscholar.org/CorpusID:168250848

[5] Pantenburg, Lukas and Nordbring, Markus. 2021. Beyond the words - A correlational study of the tone of CEO letters in relation to the stock market performance in Japan, Sweden, and the United States. https://lup.lub.lu.se/student-papers/search/publication/9046987

[6] Raghuram Rajan, Pietro Ramella, and Luigi Zingales. 2023. *What Purpose Do Corporations Purport? Evidence from Letters to Shareholders.* Working Paper 31054. National Bureau of Economic Research. https://doi.org/10.3386/w31054

[7] Ozcan Saritas, Pavel Bakhtin, Ilya Kuzminov, and Elena Khabirova. 2021. Big data augmentated business trend identification: the case of mobile commerce. *Scientometrics* 126, 2 (Jan. 2021), 1553–1579.

[8] Ahmed. S. Wafi, Hassan Hassan, and Adel Mabrouk. 2015. Fundamental Analysis Models in Financial Markets – Review Study. *Procedia Economics and Finance* 30 (2015), 939–947. https://doi.org/10.1016/S2212-5671(15)01344-1 IISES 3rd and 4th Economics and Finance Conference.