StockLens: Understanding Corporate Dynamics

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 check-in slides and presentation given on 14 November 2024. This document elaborates restates the original details of the project and provides updates regarding what work has been completed and what work is to be completed for the course of the project.

ACM Reference Format:
Peter Ly, Vijay Kumar Poloju, Harrison Fagan, and Andrew Connell. 2018. StockLens: Understanding Corporate Dynamics. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 6 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.

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

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

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 (specifically, keyword analysis, topic modelling, 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 solely on the tone of CEO letters; instead, it will concentrate on the revenue and general technical direction of the companies in addition to the tone of the letters to the shareholders.

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. Additionally, we are interested in determining commonalities between companies based on their stated goals and determining whether these commonalities give

better classifications of companies than standard classifications such as technology, healthcare, and energy.

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 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. We plan to instead compute cross-correlation between stock prices to estimate the correlation as a function of offset and window size.

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. We then specify our proposed methodology for each problem and our proposed evaluation rules for each method.

3.1 Measurement of Impact of Business Decisions on Stock Performance

We will analyze SEC 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 SEC reports and stock price data. To measure the impact of business decisions on stock performance we will do three things:

(1) For each SEC report, analyze the report for key business directions and general sentiment.
(2) For each SEC report period, determine how the stock price of the company changed.
(3) Construct three models which will output predictions for how the stock price will change. Namely, the models we intend to construct are decision tree, random forest, and LSTM model.

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 evaluate the accuracy of our models, we will separate the data into a training set and a testing set, and then compute the accuracy of the models trained on the training set on the testing set.

3.2 Identification of Patterns and Trends across Multiple Companies

We will analyze SEC reports to determine specific patterns and trends in various company's claimed directions and identify natural groupings of companies based on corporate direction. To conduct this analysis, we will need the public SEC filings of the companies

which we want to analyze. To identify the patterns, trends, and natural groupings, we will do two things:

(1) For each SEC report, analyze the report for key business directions and general sentiment.
(2) For each company-year pairing (corresponding to an SEC report), we will generate a vector encoding the company, year, key business directions, and sentiment.

After doing these two things, we will apply a clustering algorithm to identify natural groupings of companies according to their business directions and sentiment for the year.

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 evaluate 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. This will allow us to manually evaluate the usefulness of the groupings. We also plan on evaluating the effectiveness of the clustering via standard measurements such as Within-Cluster-Sum-of-Squares (WCSS) and Silhouette score.

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. To perform the correlation analysis, we will compute cross-correlation between pairs of stocks for varying lengths of time. Then, for pairs of stocks and lengths of times in which we find abnormally high correlation, we will then verify whether a partnership exists or not.

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 analysis, we will perform the analysis on a set of companies, and then compute the confusion matrix for our analysis. We are planning to automate this process, so that we can then compute accuracy of our analysis for relatively large collections of companies.

3.4 Future Stock Performance Prediction

We will analyze SEC reports and historical stock price data to generate a model which can output predictions about how the price of a stock is expected to change. Specifically, we will do two things:

(1) For each SEC report (and therefore year-company pair), we will analyze the report for key business directions and general sentiment.
(2) For each prediction period (currently we are planning to generate a prediction for each month), construct an encoding for both the key business directions and recent performance

We will then input test several models for these data such as decision trees, random forests, and long short-term memory (LSTM) models.

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 models, we will separate the data into a training set and a testing set, and then compute the accuracy of the models trained on the training set on the testing set.

4 Project progress

In this section we provide updates about our project progress.

4.1 Milestones

In this section, we give a high-level overview of the different phases of our project. We are currently cleaning data and performing initial validation of our techniques.

4.1.1 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.

4.1.2 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.

4.1.3 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.

4.1.4 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.

4.1.5 Complete modeling. 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.

4.2 Completed work

In this section, we provide a detailed overview of the work done for the project and for which phase of the project the work is used.

4.2.1 Data Collection. We acquired the data from the official SEC (Securities and Exchange Commission) website, sec.gov. This source contains comprehensive stock-related details, including financial statements of all companies spanning more than 15 years.

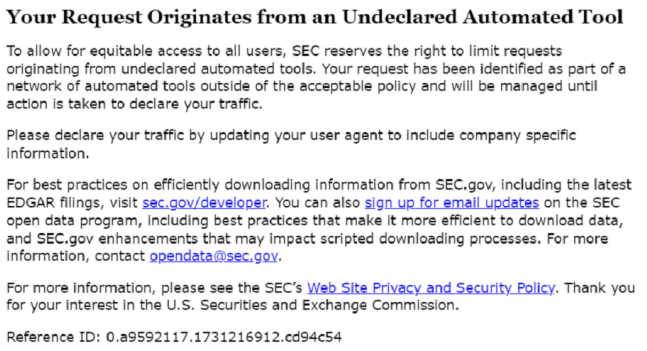
To efficiently gather this large volume of data, Python automation tools were developed and implemented. Our code utilized multithreading with workers to expedite the process, enabling simultaneous downloading of multiple files. This approach allowed us to download approximately 30GB of data in under 8 minutes.

Additionally, we bypassed the SEC's automation restrictions by identifying a loophole, allowing the use of multiple workers

```python
def download_sec_files(download_directory,  max_files=50):
    if not os.path.exists(download_directory):
        os.makedirs(download_directory)
    chrome_options = webdriver.ChromeOptions()
    prefs = {'download.default_directory': download_directory}
    chrome_options.add_experimental_option('prefs', prefs)
    try:
        driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()), options=chrome_options)
    except Exception as e:
        print(f"Error initializing WebDriver: {e}")
        return
    try:
        driver.get(website_url)
        download_links = driver.find_elements(By.XPATH, "//a[contains(@href, '.zip')]")
        download_links = download_links[:max_files]  # Limit to max_files
        file_urls = [link.get_attribute('href') for link in download_links]
        with ThreadPoolExecutor(max_workers=5) as executor:
            futures = {executor.submit(download_file, url, idx, max_files): idx for idx, url in enumerate(file_urls)}
            for future in as_completed(futures):
                future.result()
    finally:
        driver.quit()

download_sec_files(download_directory, max_files=55)
```

Figure 1: Python was used to download archives of data from the SEC website.



Your Request Originates from an Undeclared Automated Tool

To allow for equitable access to all users, SEC reserves the right to limit requests originating from undeclared automated tools. Your request has been identified as part of a network of automated tools outside of the acceptable policy and will be managed until action is taken to declare your traffic.

Please declare your traffic by updating your user agent to include company specific information.

For best practices on efficiently downloading information from SEC.gov, including the latest EDGAR filings, visit sec.gov/developer. You can also sign up for email updates on the SEC open data program, including best practices that make it more efficient to download data, and SEC.gov enhancements that may impact scripted downloading processes. For more information, contact opendata@sec.gov.

For more information, please see the SEC's Web Site Privacy and Security Policy. Thank you for your interest in the U.S. Securities and Exchange Commission.

Reference ID: 0.a9592117.1731216912.cd94c54

Figure 2: The SEC is has relatively strict guidelines regarding programmatic access of their data.

simultaneously (see figure 1 for the code and figure 2 for the SEC warning).

Post-download, custom software was written to:

- Unzip Files: Extract all contents from compressed files,
- Parse Data: Convert the extracted data into a readable format.

One major challenge was that the SEC website restricts automated tools, making it initially difficult to retrieve data programmatically. It took considerable time and effort to identify loopholes in the website's restrictions to enable the use of worker APIs for seamless data extraction from the servers.

4.2.2 Data Integration. The processed data was structured into a relational database format with the following columns:

- Tag: Represents the key or field from the SEC reports,
- Type: Indicates the type of the report, such as 10-Q or 8-K,
- Value: Contains the corresponding value of the tag,
- Time: Reflects the quarter in which the report was generated,
- Name: Specifies the company name.

To store the data, we used SQLite, a no-cost SQL database hosted in our College OneDrive. This database efficiently stored all processed data in the above relational format.

Initially, the data spanned 126GB across multiple companies. After processing, the dataset was condensed to approximately 8,000 records per company, significantly reducing storage requirements while retaining relevant insights.

The acquisition and integration process included data for the following three companies which will be used for our analysis:

- Apple Inc.,
- Automatic Data Processing (ADP),
- NVIDIA Corporation.

The programs were optimized to process each company's data in under 5 minutes, even in the worst-case scenario. This high efficiency was achieved through advanced preprocessing techniques and multithreading.

4.3 Remaining work

In this section, we provide a detailed overview of the work yet to be done for the project and for which phase of the project the work is used.

4.3.1 Data Cleaning. In the process of cleaning the database pulled from the SEC website, our first priority is to remove a large number of useless tags. There are about 521 distinct tags per company. After removing the redundant tags, we will then need to clean the information stored in the value column, which contains the actual data. This involves removing unnecessary text and formatting the remaining data into a time series structure for better analysis.

The cleaning requirements differ based on the type of data. Tags need to be evaluated and cleaned based on whether they point to important information in the value column. Meanwhile, the value column must be cleaned to extract meaningful data points and ensure consistency in format. This differentiation is crucial for ensuring the overall quality of the dataset.

The collected data is full of junk data that bloats the dataset and complicates analysis. Retaining irrelevant information not only wastes time during processing but also detracts from the focus on critical data points. By cleaning this data, we aim to improve computation efficiency significantly while minimizing the loss of information for accurate predictions.

Currently, our cleaning process for tags involves a manual review, using specific keywords to identify and eliminate irrelevant entries. This method is labor-intensive but necessary because we could not think of an automated way to clean it. For the value column, we plan to implement some form of keyword analysis to extract dates and other important information. The keyword analysis part will be automated, so this model can scale to all companies given enough computation.

After cleaning, we aim to retain key information such as dates, revenue figures, and indicators of a company's direction. Anything that will help us to make better predictions. We might add or remove some data if the model isn't coming out quite right.

4.3.2 Keyword Analysis and Topic Modeling. In our project, we are using topic modeling and keyword analysis to analyze publicly disclosed reports from the SEC, with the goal of identifying key themes and trends related to future directions companies are pursuing. Keyword analysis helps identify important terms within these documents, while topic modeling groups words that collectively represent the broader themes or topics of the reports. The primary focus of this analysis is to detect emerging technologies like cloud computing and generative AI, as well as more general concerns such as sustainability, diversity, and corporate risk management.

To conduct this analysis, we are utilizing different methods for keyword analysis and topic modeling. For keyword analysis, we are employing term-frequency inverse-document-frequency (TF-IDF), which highlights important terms by weighing their frequency against how common they are across all documents. This approach is preferred over basic term frequency, which could be skewed by commonly occurring but less meaningful words. For topic modeling, we are considering two techniques: Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA). NMF is simpler and faster to compute, but the resulting topics can be more challenging to interpret due to the linear algebraic nature of the matrix decomposition. In contrast, LDA generally produces more interpretable topics, though it is computationally more intensive. We include sample runs for TF-IDF, NMF, and LDA in the appendix.

While these methods offer powerful tools for understanding trends in the data, we have encountered challenges, particularly in setting appropriate thresholds for filtering keywords and topics. Some common words, which appear frequently across the documents, are difficult to eliminate, leading to potential noise in the analysis. Additionally, some of the topics produced by our modeling techniques do not always align with clear, coherent themes, making interpretation of topics and company trends more difficult. To that end, we are continuing to refine our approach to better capture the evolving directions that companies are taking.

4.3.3 Cross-correlation Analysis. In our project, we are using cross-correlation as a statistical tool to analyze the relationship between stock prices. Cross-correlation measures the degree to which the movements of two time series, in this case, stock prices, are related over time. This technique will allow us to assess how closely the price changes of different stocks are linked, specifically focusing on pairs of stocks within our dataset.

We plan to use cross-correlation to evaluate how closely the price changes of stocks in companies with potential partnerships are related. We expect to observe that companies with existing partnerships will exhibit a higher degree of correlation in their price movements, reflecting a stronger relationship between the stocks. However, we recognize certain limitations in applying this method. One key challenge is that some partnerships between companies may be short-lived, meaning that the cross-correlation analysis might need to be performed on a smaller, more specific time window to accurately capture these temporary relationships. This consideration is important for ensuring the analysis reflects the true dynamics of the partnerships being studied.

4.3.4 Sentiment Analysis. Sentiment analysis is a process in which language is converted into a machine decipherable format usually a number that corresponds to how positive or negative the sentence is, but sentiment analysis is a broad research area and there is no one agreed upon technique that is best. There are ML approaches as well as more traditional approaches that map words to positive or negative values.

In our project sentiment analysis will play a valuable role in identifying whether sentences within SEC filings an overall positive or negative outlook. It is a reasonable assumption from there that sentences with a more positive outlook correspond to a more positive outlook for that company, and vice versa.

As mentioned in the previous paragraph we will be using SEC filings with our sentiment analysis tool. The aggregated dataset which we have collected contains all the text within these reports that can provide a picture of the overall sentiment of the sentence. Combining this with every piece of text from the filing, we can achieve an accurate and comprehensive sentiment from the whole document. Using the whole document will allow sentences that are outliers in their sentiment as it relates to the rest of the document to not make an adverse impact in our modeling or analysis.

We expect that overall document sentiment is highly correlated with stock performance, e.g. a company with a mediocre sentiment has an overall flat stock price not seeing much movement.

For sentiment analysis there are several tools each with their own theoretical benefits, our team has chosen two tools to explore these pros and cons with, VADER and FINBERT. VADER is a more traditional approach to sentiment analysis using human graded sentiments and their algorithm. FINBERT is a machine learning model using BERT embeddings fine-tuned on financial data with a sentiment analysis aspect on top. It is trained on a sentiment data set presumably from human graders.

VADER is a highly optimized algorithm that can allow us to process more data at a higher speed giving us the ability to run this tool on more companies giving us more data for our analysis. VADER is a standard tool and is known for its robustness, but it is not tuned for finance specifically. FINBERT on the other hand is a more modern approach using state-of-the-art techniques and could give us more accurate sentiments, but being around seven billion parameters it is highly likely that inference would take much longer, giving us less companies to work with.

Sentiment analysis is a difficult task, and with the complex nature of language these tools will often misunderstand what certain words mean in context. This includes within sentences, but considering paragraphs and the whole document words can be misinterpreted easily. Still we believe that the average sentiment of the document is still accurate.

4.3.5 Case Study Validation. To validate our overall methodology for each of our four goals, we will run all of our methods on a small collection of six companies. This will allow us to ensure that our full project pipeline is working as intended and give a benchmark for the throughput of our analytic methods.

4.3.6 Complete Modeling. Finally, for the complete modeling, we intend to test our analysis on a larger set of companies from different market sectors. Because we have not been able to test our methodology on a small set of companies to determine how long our pipeline will take to run, we have have not yet determined the number of companies which we plan to analyze for this project.

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.

A Sample Keyword Analysis and Topic Modeling

We initially tested our keyword analysis and topic modeling on the following bank of sentences:

- "The stock market is volatile, and investors are looking for safer investments in uncertain times.",
- "Artificial intelligence is transforming many industries, including healthcare and finance.",
- "The latest smartphone models have cutting-edge technology, with improved cameras and faster processors.",
- "The healthcare industry is facing challenges, including rising costs and the need for more efficient care delivery.",
- "Investing in stocks can be risky, but the potential returns make it attractive to many investors.",
- "Machine learning algorithms are now being applied in various fields, from medicine to autonomous vehicles.",
- "The future of finance looks promising, with blockchain technology and cryptocurrency changing the landscape.",
- "Healthcare companies are focusing on patient-centered care and reducing operational inefficiencies.",
- "The development of new technologies like 5G and AI are revolutionizing the tech industry.",
- "Cryptocurrency and blockchain are disrupting traditional financial systems, with Bitcoin leading the way.",
- "Telemedicine has gained popularity, especially during the COVID-19 pandemic, as healthcare services shift online.",
- "Investors should be cautious with their portfolios, considering both market trends and economic factors.",
- "The tech industry continues to grow with innovations in virtual reality, robotics, and cloud computing."

From this bank we were able to extract the following:

- Keywords: artificial, artificial intelligence, care reducing, companies, companies focusing, focusing, focusing care, healthcare companies, healthcare finance, including healthcare, industries, industries including, inefficiencies, intelligence, intelligence transforming, many industries, operational, operational inefficiencies, reducing, reducing operational, transforming, transforming many, blockchain technology, cautious, cautious portfolios, changing.
- Topics:
  - Topic 1: healthcare, including, care, industries, intelligence, artificial, transforming, patientcentered, reducing, focusing
  - Topic 2: investors, market, cautious, trends, portfolios, considering, economic, factors, volatile, stock
  - Topic 3: blockchain, cryptocurrency, technology, landscape, changing, future, looks, promising, finance, systems
  - Topic 4: tech, industry, new, technologies, revolutionizing, development, ai, like, 5g, grow

We also obtained a topic clustering for the input sentences (see figure 3):

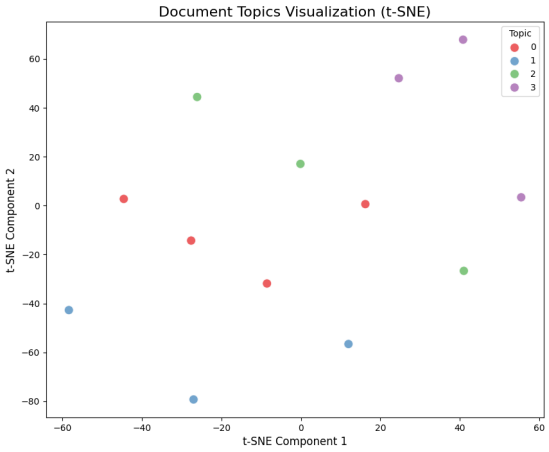


Figure 3: Each sentence is assigned a primary topic. The clusters computed here are via t-SNE (t-distributed Stochastic Neighbor Embedding), to reduce the number of dimensions to two.