Machine Learning Capstone - Final Report

Nikhila Kulukuru (nkulukur), Tony Ye (zhenghay), Qirong Wang (willwang),
Haoshu Lyu (haoshul), Sylvia Zhu (yongshiz)

November 2023

Abstract

The 10-K and 10-Q filings submitted to the SEC by all publicly traded U.S. companies have a huge amount of information. Many valuable insights can be drawn from these filings. However, due to the size of the filings and the comprehensive nature of these filings, it is very hard to draw insightful conclusions. In this paper, we aim to construct a vector database that stores textual information in numerical form. This numerical vector for every firm is calculated using various Natural Language Processing (NLP) techniques. This paper also aims to find companies that are the most similar to each other which can give businesses valuable insights into investment strategies. The other topics that this paper touches upon include classifying companies based on the GICS code and identifying profit pools within industries. The report successfully demonstrates the utility of NLP in financial analysis, offering new perspectives for strategic investment and company assessment.

1 Introduction

Publicly traded companies in the United States are mandated by the Securities and Exchange Commission (SEC) to disclose periodic financial reports, such as Form 10-K and Form 10-Q. These essential documents offer not only critical financial data but also valuable insights into a company's business model, growth strategy, and potential risks. However, the comprehensive nature of these filings can make them daunting to sift through. This research project aims to convert this wealth of textual and numerical information into a more accessible format by creating a comprehensive vector database for U.S. companies. In addition, the study intends to harness the capabilities of Natural Language Processing (NLP) techniques and Language Learning Models to construct a language model capable of answering queries related to specific sectors, companies, or clauses within the SEC filings.

2 Objectives

The project is designed to accomplish the following objectives -

- Vector Database Construction: The first step is to construct a robust and comprehensive vector database that incorporates both key financial indicators and textual content from SEC filings. Specifically, the database will focus on sections of the 10-K and 10-Q filings like Risk Factors, Management's Discussion and Analysis (MD&A), and Business Descriptions. This vector database will serve as the first step towards building more complicated language learning models which can then be used to assess companies across various industries.
- Similarity Search Language Model: The second step is the utilization of NLP techniques to create a sophisticated language model capable of analyzing the vast textual content found in SEC filings.

The model will quantify sentiment and assess how different companies relate to one another, especially within supply chains. By identifying such correlations, the project aims to offer deeper insights that could be crucial for investment strategies and risk assessment. To find the companies that are similar to each other we closely follow the cosine similarity methods highlighted in [1].

- GICS & Company Profit Pools: Another objective of this research project is to employ NLP algorithms to identify notably profitable companies within specific industries in the U.S. market. The aim is to go beyond mere profitability metrics to understand each company's role and leadership within broader industry trends and competitive landscapes. This will be achieved by scrutinizing targeted sections of SEC filings that focus on profitability, market share, and competitive positioning. The insights gained will serve to guide investment strategies and business decisions.

3 Background

As per the introduction, the first objective of this project is to create a vector database with an embedding of all the companies. In the context of NLP, the process of embedding is a technique where words or sentences are represented as a dense set of numbers. Traditional embedding methods are the bag-of-words representation and the Term Frequency-Inverse Document Frequency (TF-IDF) representation.

In our project, the data are the appended 10-K and 10-Q filings over two years for every company in the S&P500. The first approach we used to create the embedding is via a co-occurrence matrix and the first step here is to construct the Pointwise Mutual Information (PMI) matrix. This holds the Pointwise Mutual Information values between pairs of terms in a corpus or dataset. It is a statistical measure used in information theory to evaluate the association between two words/tokens/terms. To do this, we first need to compute the frequency of skip-grams, which are just generalizations of n-grams that allow tokens to be skipped within some specified window. It is equivalent to computing the frequency of times a pair of words are observed together within some defined window indicating how far apart the tokens are allowed at max to be apart from each other. After constructing the PMI matrix, we will perform singular value decomposition (SVD) on it to extract the word embeddings.

In addition, we also use the TF-IDF method of representation of the data. TF-IDF evaluates the importance of a word in a document relative to the entire set of documents. It essentially captures how unique each word is across documents. If a word has appeared in many documents, then it is considered less significant in the TF-IDF realm.

The result of the above work is a vector representation for each company. The next objective is to measure company proximity, done here via cosine similarities. This metric assesses how closely related one company is to another in a multidimensional space. This metric has the potential to unveil intricate patterns, correlations, and insights within the data, ultimately enabling more informed decision-making and strategic planning. Different conversion approaches from text to data are crucial in getting a satisfactory result for the cosine similarities of companies. This work will pursue using all two approaches - the TF-IDF matrix and the co-occurrence matrix.

After establishing the vector representation of companies based on their textual data, the next challenge lies in classifying these entities with respect to the Global Industry Classification Standard (GICS) codes. The GICS structure, designed to systematically categorize companies into distinct sectors and industries, becomes an essential reference point for understanding company profiles, benchmarking, and

portfolio construction. However, manual assignment of these codes is both time-consuming and prone to inconsistencies, especially when working with a vast array of companies like those in the S&P 500. Thus, we leverage the semantic nuances captured in embeddings to train classification models that can discern the differences between industry sectors and accurately assign GICS codes. Furthermore, the automated classification of companies into GICS codes, based on our models, may present discrepancies when compared to their existing allocations. It is essential to investigate disagreements between the model's predictions and the actual GICS codes. By doing so, we aim to discern the accuracy of our approach and identify any potential reasons for divergences.

Using the GICS industry codes from our classification models, we create pools of firms in the same industry. A profit pool refers to the total profits earned by all firms in a particular industry or market. The concept of a profit pool is often used in strategic management and business analysis to understand the overall profitability of an industry and the position of a firm in a certain industry. We try to create graphs with various metrics such as profits, profit margins, net income, etc to understand where exactly a firm stands in an industry based on its financial health. Analyzing the profit pool helps businesses and investors understand where the money is being made within a particular market.

4 Data Utilized

The data utilized in our project are quarterly and annual financial reports (10Q and 10K) published by publicly traded companies. We currently focus on all S&P 500 companies with reports from 2021 to 2023. In total, we collected 4,461 filings spanning three years from the SEC website using data scraping techniques in Python. A notable challenge we encountered during the data scraping process was the SEC website's limitation, allowing only 100 requests per session. To maximize efficiency and overcome this constraint, we implemented a systematic approach by creating batches of 100 filings for each scraping session.

For our analysis, we looked at specific parts of these reports, such as the business description, Management's Discussion and Analysis (MD&A), and risk factors. The business description section provides an essential snapshot of a company's core operations, helping us understand the business; the MD&A section offers insights directly from the company's management about its financial performance and future prospects; the risk factors section contains potential challenges and risks that might impact a company's performance. These sections were selected because they contain fundamental information about each company's business, financial health, and potential risks, forming the cornerstone of our analysis.

In addition, one of our goals is to classify each company into an industry. We obtained the GICS code from Bloomberg for each company in the S&P 500 and used them as the response variable for the industry classification problem. Another goal is to find the profit pools in an industry. For this, we download the profit, profit margin, and net income values for each company in the S&P 500 from Yahoo Finance using the yfinance package.

5 Methods

5.1 Numeric Representation & Word Embeddings of the Textual Data

The first step to be done before we can reach any conclusions is to create a numeric representation of the 10-K and 10-Q filings. One can use various word embedding techniques to represent the corpus of data. In this project, we follow the 2 methods to represent the text.

Method 1 - SVD on co-occurrence matrix

- First, we need to compute the frequency of skip-grams, which are just generalizations of n-grams that allow tokens to be skipped within some specified window. This is similar to computing the frequency of times a pair of words are observed together within some defined window indicating how far apart the tokens are allowed to be. After that, we obtain the frequency for both words in each of the word pairs and calculate PMI values for each pair using the formula -

$$\text{PMI} = \max\left(0, \log\left(\frac{\text{freq\_skip\_gram}}{\text{freq1} \times \text{freq2}}\right)\right)$$

- Extract the embeddings by doing SVD on the PMI matrix obtained in Step1 (using TruncatedSVD in sklearn with the number of components being 500).
- After doing the SVD operation, we reduce the dimension of the original skip-grams matrix into a 41092 by 500 dimensions, with all the terms on the rows and the 500 reduced dimensions on the columns. After this step, we sum embedding values across terms in documents - basically computing the average embedding vector value across the tokens in the document. This enables us to get a final matrix that has all 496 companies on the rows and the 500 reduced dimensions on the columns.

Method 2 - SVD on the TF-IDF matrix

- The TF-IDF values are good ways to compare text across documents and can help us understand which words are frequently used in one set of texts but not the other. The calculation of the TF-IDF matrix is as follows. First, we get the IDF value: for word v, compute

$$\text{idf}_v = \log\left(\frac{N}{N_v}\right)$$

  , where N is the number of documents and N_v is the number of documents with the word v.

- Then, we compute

$$\text{TF-IDF}_{i,v} = \text{tf}_{i,v} \times \text{idf}_v$$

  , where $\text{tf}_{i,v}$ is the weight of term v in document i and idf_v is the IDF value from above.

- In Python, this could be done using the TfidfVectorizer function provided by sklearn. Likewise, extract the embeddings by doing SVD on the TF-IDF matrix obtained. One caveat here is the choice of how many dimensions to reduce the original matrix into. Lastly, in the same way as Method 1, we perform cosine similarity to find similar companies.

5.2 Company Similarity

After we have the word embeddings of the textual data, our first goal is to find the top n similar companies for a particular company. We try to find the similarity between companies through the concept

of cosine similarity.

Cosine similarity is a metric used to measure how similar 2 vectors are in a multidimensional space. We can apply the formula of cosine similarity between every 2 companies to find the top n similar companies to a particular company.

The cosine similarity between 2 vectors A and B mathematically is defined as the cosine of the angle of between them. It is calculated using the dot product of the vectors divided by the product of their magnitudes.

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

, where A is the vector for the first company and B is the vector for the second company. In Python, we can calculate cosine similarity between 2 vectors using numpy and sklearn packages. The cosine similarities between two vectors can inform about the angle between two vectors(i.e. two companies) in a high-dimensional space. The smaller the angle, the closer or more similar the companies are.

5.3 GICS Classification

The GICS classification stands for Global Industry Classification Standard. Apart from finding the most similar companies, we also want to see if the GICS classification for all firms is accurate. Hence, we build a Random Forest Multiclassification to see whether companies fall into the same industry as their GICS classification. We also try to use bagging with random forest and gradient boosting to improve the accuracy of classification and use grid search for parameter tuning.

For companies that do not fall into the same industry, we look deeply into the company's composition to understand if they might be better suited for another industry. The main goal of this section is to see whether the GICS classification has missed anything important about firms in a certain industry. The classification results could be further used in the profit pools analysis.

5.4 Profit Pools

The primary objective of GICS classification is to create profit pools across different industries. To accomplish this goal, our approach involves obtaining up-to-date financial data, specifically profit and profit margin figures, from Yahoo Finance for each company. Subsequently, we utilize this data to construct interactive graphs illustrating the relationship between profit margins and financial metrics, such as Net Income, spanning three years. This analysis allows us to identify companies that significantly contribute to the profit margin within their respective industries. Companies can use this knowledge to identify attractive market segments, allocate resources effectively, and develop competitive strategies to capture a larger share of the profits.

6 Results

6.1 Company Similarity

The overall results of Method 1 are satisfactory. However, at the start, the result based on the raw SVD embedding is not satisfactory, and this may be because there are many common words such as 'finance',

'statements', 'economy', etc, appearing across different documents. These words might make different companies' documents look similar to each other, but after removing those words, the results are much more satisfactory. The rationale here is to remove words based on word count and remove those words that appear in a given threshold number of documents. We illustrate the results of Method 1 with some examples.

Table 1: Method 1 - Companies Similar to AAPL

Ticker	Cos Similarity
MTCH	0.839499
NFLX	0.834295
GOOGL	0.812331
GOOG	0.806822

As expected, Match Group (MTCH) is an American internet and technology company headquartered in Dallas, Texas. It owns and operates the largest global portfolio of popular online dating services including Tinder, Match.com, Meetic, OkCupid, Hinge, Plenty of Fish, OurTime, and other dating global brands. Netflix (NFLX) is an American subscription video-on-demand over-the-top streaming service. The service primarily distributes original and acquired films and television shows from various genres, and it is available internationally in multiple languages. Last but not least, the third and fourth most similar companies are both Google, with a caveat that the third most similar company is with the ticker GOOGL (the stock that has voting rights). Similarly, the fourth most similar company is with the ticker GOOG (the stock that does not have voting rights).

Below, we illustrate the method 1 results for a few more companies - AIG (American International Group), AMD (Advanced Micro Devices), and DE (Deere & Company).

Table 2: Method 1 - Companies Similar to AIG

Ticker	Cos Similarity
LNC	0.930712
MET	0.928433
HIG	0.924095
AMP	0.913994

Table 3: Method 1 - Companies Similar to AMD

Ticker	Cos Similarity
NVDA	0.801940
PH	0.716056
TXN	0.715260
HPQ	0.711216

Table 4: Method 1 - Companies Similar to DE

Ticker	Cos Similarity
J	0.554110
CAT	0.549997
TRMB	0.543007
TSCO	0.541800

The result for Method 2 is less promising. We used the TF-IDF matrix which has different companies as the rows and the different terms as the columns. We then performed SVD on this document-term matrix.

After reducing the dimensions to 200, we get a final matrix of the shape 496 by 200. We illustrate the results of Method 2 with some examples.

Table 5: Method 2 - Companies Similar to AAPL

Ticker	Cos Similarity
KMI	0.543170
PODD	0.450306
TXN	0.372224
CPT	0.323344

Again, to illustrate this result using an example, we would like to find out the top 5 companies to AAPL. Kinder Morgan (KMI) is one of the largest energy infrastructure companies in North America. The company specializes in owning and controlling oil and gas pipelines and terminals. C.H. Robinson Worldwide (CHRW) is an American transportation company that includes third-party logistics. The company offers freight transportation, transportation management, brokerage, and warehousing. It offers truckload, less than truckload, air freight, intermodal, and ocean transportation. Texas Instruments (TXN) Incorporated is an American technology company headquartered in Dallas, Texas, that designs and manufactures semiconductors and various integrated circuits. Lastly, Insulet Corporation (PODD) is an innovative medical device company dedicated to making the lives of people with diabetes and other conditions easier through the use of its Omnipod product platform. Intuitively, these companies are less similar to AAPL, even though they might have some business lines similar.

Below, we illustrate the method 2 results for the following companies again - AIG (American International Group), AMD (Advanced Micro Devices), and DE (Deere & Company).

Table 6: Method 2 - Companies Similar to AIG

Ticker	Cos Similarity
AOS	0.854100
PAYC	0.814381
OMC	0.697544
TRV	0.131939

Table 7: Method 2 - Companies Similar to AMD

Ticker	Cos Similarity
PH	0.397125
NVDA	0.244966
LIN	0.186502
CAT	0.109700

Table 8: Method 2 - Companies Similar to DE

Ticker	Cos Similarity
SHW	0.120662
URI	0.091651
BRK.B	0.052336
TTWO	0.048547

Overall, the strength of co-occurrence is very clear over the use of TF-IDF. Co-occurrence captures the context of words. This means that the words that are used in similar contexts are considered closer in the

high dimensional space. Closer here again means the angle between two vectors in the high dimensional space. The capture of contextual meaning yields co-occurrence matrix power in determining similar companies. The use of SVD after getting the co-occurrence matrix is also very useful. It may be that we have several variables that are very highly correlated, e.g., when they are heavily influenced by a small number of underlying factors, and we wish to recover some approximation to the underlying factors. Hence, the combination of applying SVD on the co-occurrence matrix produced good results.

6.2 GICS Classification

We explore the predictability of using TF-IDF as input for the GICS classification model. In this study, we divided the data into 75% for training and 25% for testing. We fitted models using logistic regression, random forest, and bagging with random forest, as well as XGBoost, and employed grid search for tuning the parameters. The results are as follows,

Table 9: Results for GICS Classification Model

Methods	Accuracy
Logistic Regression	12.1%
Random Forest	71.77%
Bagging with Random Forest	63.71%
Gradient Boosting	72.58 %
XGBoost	75%

We can see from the results that the Logistic Regression model exhibited notably low accuracy at 12.1%, potentially due to its linear approach which may not effectively capture the intricate patterns present in TF-IDF data. In contrast, ensemble methods like Random Forest and Gradient Boosting showed significantly better performance with accuracies of 71.77% and 72.58%, respectively, highlighting their strength in handling complex data structures. Interestingly, Bagging with Random Forest resulted in a lower accuracy of 63.71%, suggesting that this particular combination might not be optimal for our dataset, possibly due to overfitting or an imbalance in model complexity and data diversity. The standout performer was XGBoost, achieving the highest accuracy of 75%. This can be attributed to its advanced regularization features, which enhance its generalization capabilities and robustness against overfitting. These results not only underscore the effectiveness of ensemble methods in complex classification tasks but also emphasize the need for careful model selection and tuning, as evidenced by the varied performances.

By analyzing the results of the GICS classification, the data reveals a mix of correct predictions and misclassifications. Notably, sectors such as 'Consumer Discretionary', 'Industrials', and 'Information Technology' frequently appear in these misclassifications, suggesting that these areas might be more complex or diverse, leading to challenges in prediction. However, upon closer examination of the predictive classification results, it becomes evident that the original GICS codes are more accurate. This indicates that the results are not entirely satisfactory and that there is room for further improvement in the models.

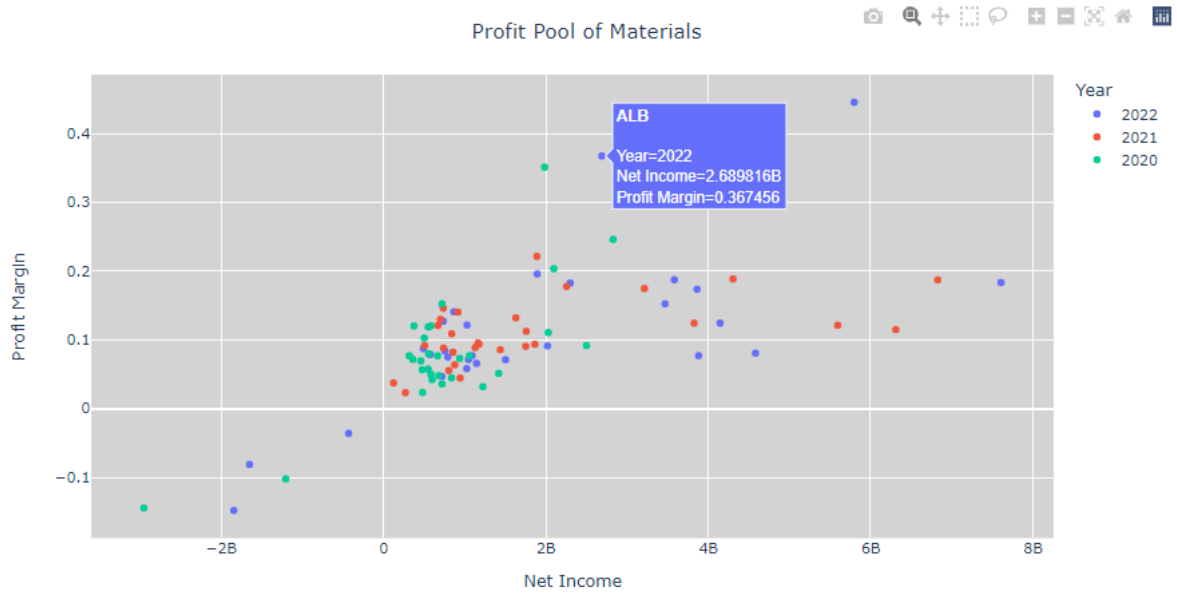
6.3 Profit Pools

Following our analysis of GICS classification, we transition to generate profit pools. In the GICS classification analysis, we concluded that the original GICS codes are more accurate. Consequently, we opted to employ these original GICS codes in the generation of profit pools. To facilitate this, we developed dynamic tools capable of visualizing diverse financial metrics extracted from multiple years of annual

statements for each industry, all presented in a single plot. For this analysis, we utilized profit margin vs. net income data from 2020 to 2022 and generated profit pool plots.

An example of the result is shown in 1 below. The plot shows the profit margin and net income of the Materials industry. Each color represents a different year, facilitating a comprehensive trend assessment over a multi-year span. Notably, the company ALB (Albemarle) stands out as a top performer. Albemarle is an American specialty chemicals manufacturing company that operates in lithium, bromine specialties, and catalysts. It is relatively unknown to the general public. However, it generated impressive profit margins exceeding 35% in 2022, positing itself among the industry's top performers. Further research reveals Albemarle as a key supplier of lithium batteries to Tesla. This analysis underscores the ability to uncover lesser-known companies with notable profit margins, such as Albemarle, thereby highlighting the value of our approach in identifying lucrative opportunities within the industry.

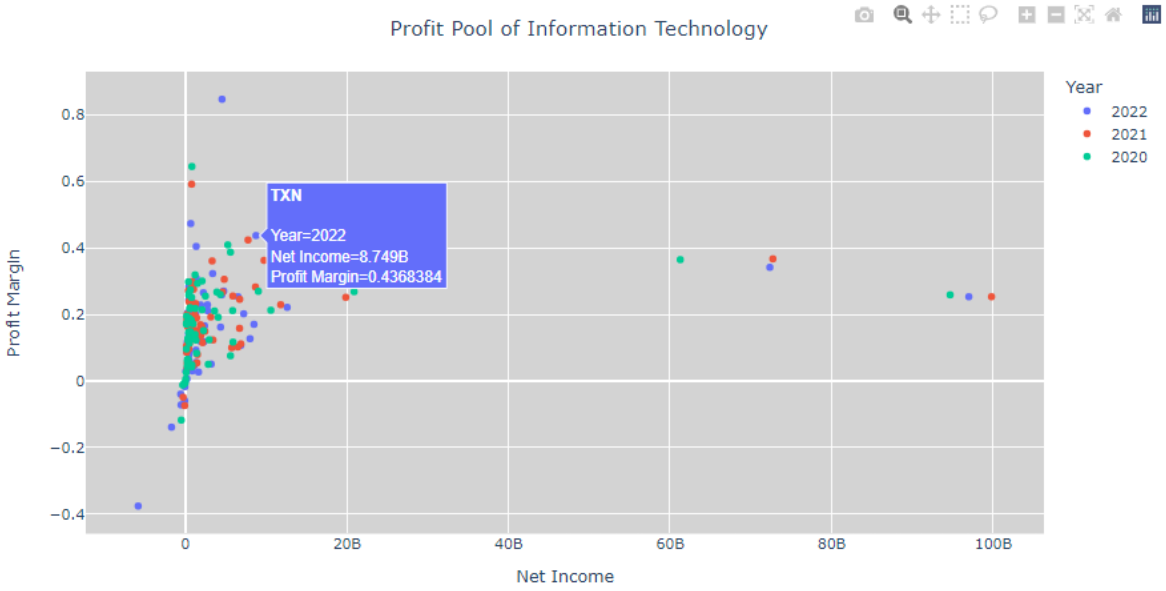
Figure 1: Profit Pool of Materials Industry from 2020 to 2022



In addition to identifying relatively unknown companies, the profit pools analysis also proves instrumental in identifying well-known companies that outshine industry peers in terms of profit margins. Figure 2 below illustrates the profit pool of the information technology sector. One noteworthy standout is TXN (Texas Instruments), generating an impressive profit margin surpassing 40%. Texas Instruments has consistently delivered robust profit margins over the past three years. This remarkable profitability is attributed to its practice of manufacturing the majority of its components in-house, mitigating profit loss to external companies. While Texas Instruments is a familiar name, this analysis unveils previously undisclosed facts about its financial strength, providing valuable insights into its substantial profit margins and robust foundation for future growth.

The tools we developed offer a high degree of flexibility, allowing users to tailor it by incorporating diverse financial metrics from annual statements and adjusting the date range as needed. Its flexibility opens up opportunities for exploring a broader array of valuable financial indicators in future analysis. Furthermore, while our current focus centers on SP 500 companies, the tool is robust enough to seamlessly expand its coverage to a more extensive range of companies, enhancing its applicability and relevance in

Figure 2: Profit Pool of Information Technology Industry from 2020 to 2022



a broader financial landscape.

7 Conclusion

This project successfully demonstrated the efficacy of various analytical methods in assessing company similarities and GICS classification, using advanced techniques like SVD, TF-IDF, and various machine learning models. The results highlighted the significant impact of word co-occurrence and contextual relevance in determining company similarities, with the co-occurrence matrix combined with SVD yielding particularly insightful outcomes. For the GICS classification, the superiority of ensemble methods like XGBoost was evident, showcasing their robustness and accuracy in handling complex datasets. The profit pools analysis further enriched our understanding, revealing both well-known and lesser-known companies as significant contributors to industry profits, emphasizing the importance of in-depth financial analysis. The flexibility and adaptability of the developed tools pave the way for broader applications, potentially encompassing a wider array of companies and financial metrics.

For future scope, this project opens up several avenues for further research and development. One potential area of expansion is the application of these methods to a larger dataset, extending beyond the S&P 500 companies to include more firms from Russell 3000 and even international corporations, which would provide a more comprehensive view of the global market dynamics. Additionally, integrating real-time data analytics could offer more up-to-date insights, crucial for investment and strategic business decisions. The incorporation of AI and machine learning algorithms for predictive analysis can also be explored to forecast market trends and company performances. Furthermore, we could use other kinds of pre-trained word embeddings like FinBert and see whether the accuracy can be increased. Overall, the project sets a solid groundwork for future explorations, offering significant potential for enhancing and broadening the scope of financial analysis and insights.

References

[1] Vamvourellis, Dimitrios, et al. "Company Similarity using Large Language Models." arXiv preprint arXiv:2308.08031, 2023

[2] https://simonwillison.net/2023/Aug/27/wordcamp-llms/retrieval-augmented-generation

[3] https://sec-api.io/docs/sec-filings-item-extraction-api/python-example