Supporting Fundamental Analysis: Combining LLMs and Machine Learning to Analyze SEC Annual Reports.

Mohamed Elmanzalawi
mye5156@psu.edu
The Pennsylvania State University
State College, Pennsylvania, USA

Figure 1: United States Securities and Exchange Commission.

ABSTRACT

In times of economic instability, the process of investing in stocks through technical analysis1 becomes a volatile task. As such, examining a company's intrinsic value through fundamental analysis2 of comprehensive financial documents becomes a more reliable investing strategy. Despite this improved investing approach, fundamental analysis remains a difficult task to implement due to the complexity and unstructured nature of financial documents such the 10-K. In this experiment we addressed the issues associated with fundamental investing through a comprehensive machine learning pipeline. This provides a supportive system to help inform investors of a company's financial health when deciding whether or not to invest in a company.

Conference'17, July 2017, Washington, DC, USA

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

KEYWORDS

Fundamental Analysis, Financial Documents, 10-K, SEC, Natural Language Processing, Word Embedding, Multi-Class Classification, Comparative Testing

ACM Reference Format:
Mohamed Elmanzalawi. 2024. Supporting Fundamental Analysis: Combining LLMs and Machine Learning to Analyze SEC Annual Reports.. In . ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

1.1 Problem

In the realm of fundamental analysis, examining financial reports that companies file annually to the Securities and Exchange (SEC) are a necessity. The most common and informative report is known as the 10-K. This report includes a variety of sections regarding a company's financial health such as business summary, properties, legal proceedings, market risk, and financial statements. Observing these sections reveals two major problems: 1) the extensive nature of the information, and 2) the amount of unstructured information. The combination of lengthy and unstructured information makes exploring the 10-K report a challenging and time consuming task.

1.2 Difficulties

Despite these difficulties, exploring and analyzing such complicated financial information remains necessary for fundamental analysis. Currently, financial experts are trained for many years before being fully capable of assessing a company from such documents. They often employ the computational based approaches that summarize and structure text to aid their analysis. But representing extensive, unstructured sections as conscious and structured forms comes with its own difficulties. Naive approaches to document reduction and text representation, such as Bag of words and TF-IDF, often generate sparse and computationally complex data.

1.3 Improved Solution

Fundamental analysis is usually a task left to certified financial professionals and investment portfolio managers. The majority of innovation in this area is stringently shielded by intellectual and property rights of investing and financial services companies. While this lack of transparency contributes to the lack of available open source methods, many individuals have undertaken to improve the process of stock selection and charred their methods and projects on the internet. In this work we aim to modify one of these existing methods which leverages large language models to parse and analyze large financial documents to answer specific financial questions.

1.4 Our Approach

Improving upon the previous analysis is a major aspect of this work. Our objective is to reduce the time and computational complexity required to reproduce such intricate analysis of 10-K financial documents. We achieve this objective through three primary improvements: The first is to improve the quality of financial queries used to generate machine learning features. The second improvement is reducing the number of generated machine learning features to reduce computational complexity. Finally, the third improvement is to utilize proper machine learning models and techniques to accommodate the modified features and maintain stock company predictive ability.

1.5 Evaluations

We will use the AI writing assistant Gramerly to evaluate current questions and modify them to improve the scores. Implementing the same analysis with something other than linear regression and comparing the results. If the reduced analysis can generate similar results then the improvements are self-evident.

2 RELATED WORKS

There are three main areas of related work that involve the use of the 10-K financial documents: evaluating stock sentiment, assessing investor risk associated with a company, and finally, predicting stock performance. We will further discuss these related areas by examining prominently associated research papers and the impact they each have on the process of fundamental analysis.

2.1 Sentiment Analysis

Let's begin with the first and most popular related area: evaluating stock sentiment. Financial documents such as the 10-K and the 8-K are a valuable source of company information that disseminated into the news and media over time. These documents contain crucial and standardized information regarding a company's pending legal actions, major mergers and acquisitions, and other mandatory corporate disclosures that fundamentally sway the media coverage and stock perception. Papers such as [3] "Stock price prediction through sentiment analysis of corporate disclosures using distributed representation" and [2] "Sentiment Analysis of 10-K Filings: An Approach to Automatic Processing of the Information Hidden in Accounting Narratives" believe in the eminence stock insights that can be derived from financial documents.

2.2 Risk Analysis

Now that we have examined the first major areas of related works, let's continue by discussing the use of financial documents to assess risk associated with investing in a company. As mentioned previously, 10-K documents contain crucial comrperate information. Additionally, These documents also contain information regarding companies' risk although less directly. Changes in a business operations, senior leadership/ management, and liquidity or debt obligations collectively impact a company's risk to investors. As this information is less apparent, the use of financial documents for risk analysis is centered around predicting abnormal stock returns. Works such as [4] "Predicting Abnormal Bank Stock Returns Using Textual Analysis of Annual Reports" require the use of advanced neural network approaches to extract insights.

2.3 Stock Prediction

Finally, let's examine the use of financial documents to improve the prediction of stock performance. This long-standing practice has experienced recent developments in Natural Language Processing and Large language Models that have significantly increased non-institution implementations. For example, [1] "Predicting Stock Performance Using 10-K Filings: A Natural Language Processing Approach Employing Convolutional Neural Networks" was able to generate major improvements to stock prediction through the use of deep learning techniques. Additionally, [9] "GPT-InvestAR: Enhancing Stock Investment Strategies through Annual Report Analysis with Large Language Models", was able to demonstrate great returns of an investing portfolio by selecting the most promising company's stocks.

As mentioned previously, our experiments are directly modified implementations of Udit Gupta's research paper. The author's novel use of large language models to create machine learning features has influenced us to recreate and improve his implementation. In the following sections we will detail our modified implementation and its effects on the project outcomes.

3 CONTRIBUTIONS

3.1 Data Collection

The data in this experiment is collected directly from the SEC online archive system, EDGAR. The process of collection is completed in two distinct steps: First is locating the appropriate 10-K documents and downloading the respective HTML files. HTML is the most compatible and readable format for the documents available on EDGAR due to the historic nature of the data. Second is the conversion of each HTML file into PDF format. This conversion is done to improve the efficiency of further processing and enhancing the collection of supportive stock data from Yahoo finance. This supportive data is processed into target features for future machine learning tasks.

3.2 Document Embedding

As mentioned previously, the now 10-K documents in PDF format are primed for document embedding. Document embedding is the process of converting text into numeric vector representations. With the help of a Sentence Transformer Model available on Hugging Face, each PDF document is processed and stored in a binary file format. These binary files are used in the next stage of the experiment to create machine learning features that become training columns in our final models.

3.3 Feature Generation

Generating the features that complete our data involves the use of prompt engineering in the form of asking questions to ChatGPT 3.5 Turbo. Our binary files are in perfect format to give to ChatGPT and ask questions about a company's financial well being. ChatGPT can aggregate and summarize information that all pertain to questions about a company's financial health. The use of questions such as: "Does the company have a clear strategy for growth and innovation? Are there any recent strategic initiatives or partnerships?" allows us to extract our desired information from the 10-K financial documents and use them to predict a company's potential for stock investing.

3.4 Predictive Analysis

After generating the entirety of the training and testing data frames, predictive analysis is performed with a simple linear regression model. The aim of this analysis is to predict the returns a stock can generate based on the engineered features. This prediction can ultimately reveal the top k stocks which if combined in a portfolio can outperform the Standard Poor (SP) Stock Index. This index is a collection of thousands of stocks and generates returns based on the average returns of all the companies within the index. Our analysis suggests a significant increase in returns when investing in the curated stocks from or predictive analysis.

4 NOVELTY

Throughout my application of the author's original work, I have needed to adopt minor changes to ensure that the code runs successfully. This in no way reflects upon the author's project quality, but rather highlights difficulties that arise when attempting to recreate work. I found that I needed to test every function before running the main function. In addition to the minor consistent modifications to the authors code, many of the data extraction steps have been modified to run in batches to increase the flexibility and accommodate the resources available to me as a student.

While such minor modifications were necessary for recreating the author's results, major modifications to assumptions and techniques were applied in my implementation of this project. These modifications include 1) Improving upon the features generated for predictive modeling and 2) Improving the breadth of the predictive analysis.

4.1 Improving Features

The performance of any machine learning task depends heavily on the training data, specifically the training features. If these training features are flawed, the foundation of our predictive analysis becomes unreliable. That is why I decided to create better questions that are guided by the SEC's investor documents() and supported by the process of fundamental analysis.

We have generated a total of 20 questions that help the purpose of examining a company's intrinsic value. Ten of which are financial related questions and ten are managerial related questions that aim to leverage Chat GPT's ability to extract valuable insights from extensive 10-K documents. Below are example of financial and managerial questions:

- Financial: "Has an independent accountant audited the company's financial statements?"
- Management: "Does the company ensure compliance with laws and regulations?"

In addition to the value of the questions, they are also grammatically concise and written in a uniform manner. Because algorithms assign higher importance to larger features, each question is framed such that a higher score is always a positive trait. For example:

- Management: "Is the company's employee turnover low? Is employee turnover lower than the industry average?"

In this question, we ask if employee turnover is moderate and if it is moderate compared to the company's industry standard. Because it is a red flag if a company has a high employee turnover we would not want to assign a higher score to this company if turnover is high. Instead we want the company to score higher if it has a low turnover ratio. All questions with negative implications are written with this in mind. More can be found about the question in the "questions-dict.json" file.

4.2 Holistic Predictive Analysis

Because the majority of the project is collecting, and processing the data, I felt as if the predictive analysis section lacked depth. The author only implemented a simple linear regression model to predict the top k stocks to invest in. Despite significantly outperforming the SP 500 index with this algorithm's prediction, I believe that more in depth predictive analysis could improve results even more. That is the second major implementation I contribute to this project.

As a student of computer science and applied machine learning, I wanted to apply techniques and pipelines that I have learned in my academic training to this project. The major thing that I have learned is to avoid using a single predictive algorithm and similarly avoid using a single evaluation metric. In this project I applied this wisdom by testing an additional Elastic Net regression and evaluating with the authors percentage returns and cumulative returns metric. The results of these modifications are discussed in detail in the following section.

5 RESULTS

When comparing linear regression and Elastic net regression, it's important to understand their strengths and weaknesses. Linear regression is a standard model used for predicting continuous variables. While it's commonly used, it's viewed as a baseline due to the algorithm's sensitivity to outliers. On the other hand, Elastic Net regression implements a regularization penalty to limit the impact of outliers on the prediction.

Because stocks can have vastly different returns, the percentage returns and cumulative returns of some companies are outliers to the data. This was the main reason behind implementing the Elastic net regression. Our results show a significant improvement in returns when trained using Elastic net Regression:

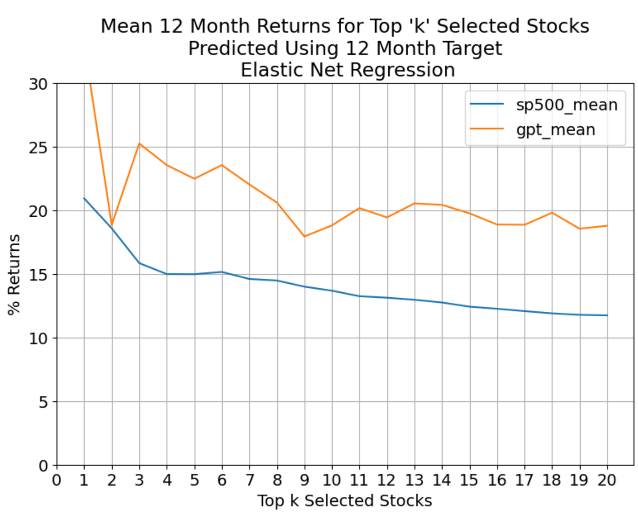

Figure 2: Elastic Net Regression Percentage Returns Results

While these returns are very promising, our experiment and implementation was very costly, from time and resource standpoints. These restrictions limit the reproducibility of our analysis and beg the question of whether improvements can be made? In the next sections, we will discuss the limitations of our experiment as well as potential improvements.

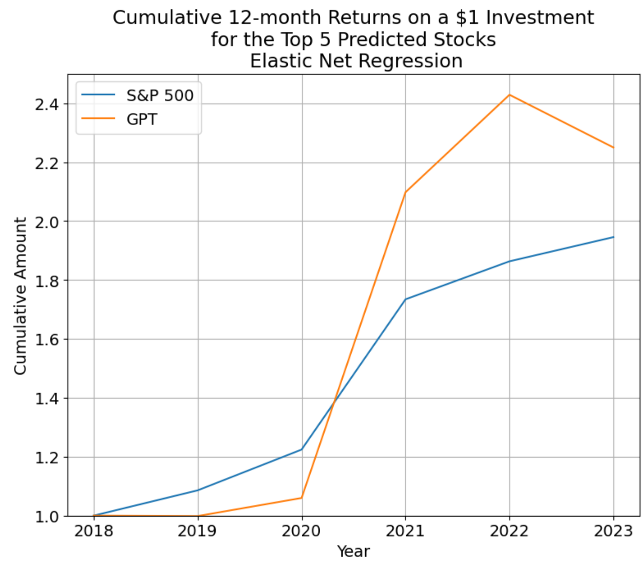

Figure 3: Elastic Net Regression Cumulative Returns Results

6 LIMITATIONS

6.1 Storage limitations

This experiment requires storage of up to 22 GB of data for the 1500 10-K documents which are downloaded as HTML files, Converted to PDF files, and finally transformed into binary files as full document embeddings. These storage needs were a limiting factor in my analysis and required the purchasing of Google Drive's 2TB cloud storage plan for 10 USD/month to supplement storage on my local computer.

6.2 Computational Power

Google Collab has been my primary code development platform. As such, I have also been required to purchase Google Collab's Plus subscription for an additional 20 USD/ month. This subscription offers 200 compute units and gives users priority to google servers so that processing of hundreds of documents may be completed in a reasonable time. Without the Plus subscription and additional compute units, running 10 document embeddings could have taken 89 hours to complete.

Even with the Google Collab Plus subscription, The maximum runtime for any Collab notebook is 24 hours which presents another limitation: time restriction. Time has been a crucial resource while recreating this analysis. Not only have runtime limits constricted and wasted time, but additional responsibilities have also contributed to the problem.

While running a Collab notebook, it's inadvisable to navigate away from the notebook and perform other tasks. Doing so could cause the runtime to discount and halt execution. This meant that physical presence was an absolute requirement to run code successfully. Unfortunately, this presence limited my ability to uphold other responsibilities as a student.

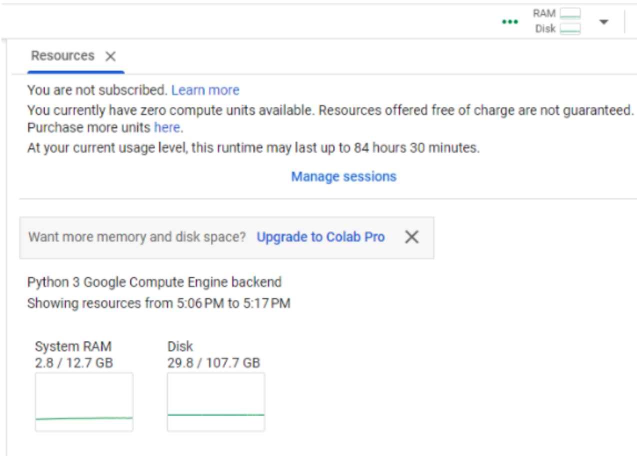

Figure 4: 10 document embeddings require more than 80 hours of runtime.

7 FUTURE WORKS

I believe there are three potential future improvements that can be done to this analysis: 1) Reduce the contents that require embedding 2) Using embeddings as features 3) Testing different transformer models. In the following sections, I will discuss the improvement, pros and cons and a possible outcome associated with the modification.

7.1 First Improvement

The first improvement is to reduce the contents of the 10-K documents in PDF format before subsequent document embedding. This can be achieved by deleting redundant and irrelevant pages within the document. This process would reduce the time complexity and memory required for document embedding. As well as reduce the tokens that are used by GPT in the feature generation step. However, it is possible to remove important information accidentally, which would limit feature quality as a result. That is why this elimination step would require extreme care to ensure no crucial information is removed. If done correctly, I believe the process of reducing the 10-K document contents will improve the time efficiency and reduce time required to process the documents.

7.2 Second Improvement

The second improvement is to directly embed common sections in all 10-k documents (such as) and use them as features, instead of using them to generate a score with Chat GPT. By embedding with the GloVe, Word2vec, Fasttext algorithms from the Gensim library, we can reduce the dependency on a GPT model to produce features for training and testing. Ultimately, this reduced dependency will reduce the cost associated with this type of fundamental analysis. Overall, I believe that using direct embeddings as features can improve predictive modeling compared to the Two-step GPT approach for generating features.

7.3 Third Improvement

Finally, Testing different transformer models can generate a more in depth comparative analysis. This can be done by generating 10-k document features with the use of Chat GPT 4.5 Turbo API or with the LLaMA 2 pre-trained model instead of GPT 3.5 Turbo. Using chat GPT 4.5 turbo can significantly reduce time required while costing more and using LLaMA 2 will not cost anything but it will require local storage compared to using Chat GPT's API. Overall I believe that Experimenting with different transformer models can create a richer comparative analysis between 4.5 turbo, 3.5 turbo, and LLaMA.

ACKNOWLEDGMENTS

Thank you to the teaching staff of DS 340 and Pennsylvania State University. A special thank you to my former partner Ahmed Mosaad who helped me in the earlier stages of this project.

8 REFERENCES

[1] Regenburg Jønsson, Kasper, and Jonas Burup Jakobsen. Predicting Stock Performance Using 10-K Filings: A Natural Language Processing Approach Employing Convolutional Neural Networks. 15 May 2018, research-api.cbs.dk/ws/portalfiles/portal/59781036/500855-Masters-thesis.pdf.

[2] Sentiment Analysis of 10-K Filings. Master's Thesis, 15 May 2018, pp. 4–4, research-api.cbs.dk/ws/portalfiles/portal/59777420/432946-Master-s-Thesis.pdf.

[3] Kim, Misuk, et al. "Stock Price Prediction Through Sentiment Analysis of Corporate Disclosures Using Distributed Representation." Intelligent Data Analysis, vol. 22, no. 6, Dec. 2018, pp. 1395–413, doi:10.3233/ida-173670.

[4] Hájek, Petr, and Jana Boháčová. "Predicting Abnormal Bank Stock Returns Using Textual Analysis of Annual Reports – a Neural Network Approach." Communications in computer and information science, 2016, pp. 67–78, doi:10.1007/978-3-319-44188-7-5.

[5] Hájek, Petr. "Combining Bag-of-words and Sentiment Features of Annual Reports to Predict Abnormal Stock Returns." Neural Computing Applications (Print), vol. 29, no. 7, Aug. 2017, pp. 343–58, doi:10.1007/s00521-017-3194-2.

[6] Indeed Editorial Team. "13 Regression Types and When to Use Them in Data Analysis." Indeed Career Guide, 4 Feb. 2023, www.indeed.com/career-advice/career-development/regression-types.

[7] SEC Office of Investor Education and Advocacy. "Investor Bulletin: How to Read a 10-K." SEC Office of Investor Education and Advocacy, 2019, www.sec.gov/files/reada10k.pdf.

[8] 15 Questions You Need to Ask About the Fundamentals of Any Stock - Motilal Oswal. www.motilaloswal.com/blog-details/15-questions-you-need-to-ask-about-the-fundamentals-of-any-stock/1758.

[9] Gupta, Udit. "GPT-InvestAR: Enhancing Stock Investment Strategies Through Annual Report Analysis With Large Language Models." arXiv.org, 6 Sept. 2023, arxiv.org/abs/2309.03079.

Received 15 April 2024