US Stock Market Analysis

Research Abstract

This study explored the feasibility of predicting corporate growth or decline using natural language processing techniques applied to annual financial reports. Employing a dataset of 342 U.S. listed companies, we conducted experiments using a combination of traditional machine learning models (Support Vector Machines, Naive Bayes) and state-of-the-art deep learning models (Recurrent Neural Networks, BERT). To evaluate model performance from an investment perspective, we introduced a novel metric: the non-fail rate, which measures the accuracy of predicting either growth or decline. Our findings indicate that ensemble methods, combining multiple well-performing models with dynamically adjusted weights, significantly improved prediction accuracy, achieving a non-fail rate of approximately 80%. However, the relatively small dataset size and the simplicity of the classification task limit the generalizability of our findings. Future research can address these limitations by expanding the dataset, exploring more granular classification schemes, and incorporating additional textual features and external factors.

I. Introduction

A. Motivation

Traditionally, mainstream research on financial performance of a company is only based on quantitative figures and financial metrics, such as gross margin, net profit or EPS. However, there is still a wealth of valuable information in the text of the annual report. The Management Discussion and Analysis Section of the annual report contains insights into future strategies, potential risks, and industry trends that quantitative data can't provide. Since Management has access to more internal information, by analyzing the MD&A, we can gain a deeper understanding of their thoughts and identify the relationship between financial performance and their usage of keywords. Take NVIDIA's annual report for example, the sentence "We made substantial strides in **broadening** our supply base to **scale** our company and better serve customer demand.". The keywords such as "broadening" and "scale" can provide positive information.

B. Objectives

The main objective of the report is to analyze the companies within the S&P 500 index and predict the direction of their financial performance metrics for the next year. We apply multiple classification and clustering models to predict financial metrics, and we examine their performance and optimize their precision. We aim to find the most effective model or combination of models that can provide accurate predictions of financial performance. Additionally, we would like to develop a stable and accurate tool for annual report analysis. The tool can enable investors to have more comprehensive information, hence enhancing investment performance for investors.

C. Literature Review

In our study, we aim to use the language of financial reports as input to predict whether economic indicators will show growth (Good), stagnation (Draw), or decline (Bad). In the research "An Exploratory Study of Stock Price Movements from Earnings Calls"[1], the authors utilized linguistic data from earnings call transcripts to predict stock performance. Their approach involved sentiment analysis using Word2Vec and applied StockGNN as the predictive model. Similarly, the work by Sourav Medya et al. highlighted that "soft" data, such as sentiment-related language, can sometimes outperform "hard" indicators in predicting future stock price changes. Their study reported precision rates of approximately 0.6 across various models, suggesting that leveraging textual data for forecasting economic indicators is a viable methodology.

Beyond Sourav Medya's research, the study by Stefan Feuerriegel et al., titled "Long-term Stock Index Forecasting Based on Text Mining of Regulatory Disclosures", provides further foundational insights for our research. In their work, sentiment analysis was conducted using the Loughran-McDonald finance-specific dictionary, and they employed models such as lasso, ridge regression, elastic net, gradient boosting, principal component regression, and random forest for analysis. While our study does not employ the same models, we also use TF-IDF vectors as input features. Stefan Feuerriegel et al. [2] also referenced the Efficient Market Hypothesis (EMH) to explain that stock valuations adjust according to new information entering the market, which includes the release of financial reports.

Given that the studies mentioned above predominantly focus on data derived from telephone conference calls or regulatory reports, we are curious whether other information sources may also impact economic indicators. The study by Bathini Sai Akash et al[3]. examined widely recognized sources such as news articles and Twitter posts. They conducted sentiment analysis on these textual data and built predictive models using the extracted sentiments as inputs. Their findings revealed that combining sentiment analysis with stock return rates achieved an accuracy exceeding 60%, particularly for predicting the Dow Jones Industrial Average (DJIA) and other companies.

[1] Medya, S., Rasoolinejad, M., Yang, Y., & Uzzi, B. (2018, June). An exploratory study of stock price movements from earnings calls. In *Proceedings of Woodstock '18* (pp. 1–XX). Woodstock, NY.

[2] Feuerriegel, S., & Gordon, J. (2022). Long-term stock index forecasting based on text mining of regulatory disclosures. *ETH Zurich* and *University of Freiburg*, Zurich, Switzerland, and Freiburg, Germany.

[3] Akash, B. S., & Dagli, C. (2022). Real-Time online stock forecasting utilizing integrated quantitative and qualitative analysis. *BITS Pilani Hyderabad Campus* and *Missouri University of Science and Technology*, 1–23.

II. Methodology

A. Data Preparation

In this report, we examine the companies in the S&P 500 Index. We exclude companies that are unsuitable for using gross margin to evaluate performance, such as the finance industry, credit card company, holding companies, and any companies whose gross margin data can’t be found. After filtering, we examine a total of 342 companies. We extract textual description in Management Discussion and Analysis sections of their annual report and exclude some unimportant sections such as accounting estimates which don’t provide financial information to our analysis. For each company, we extract about 2000 to 5000 words of texts from the MD&A.

B. Label Selection

We choose two financial metrics as our objective indicators. One is the growth of gross margin rate, and the other is the growth of Return on Assets (ROA). Gross margin rate is an indicator that evaluates how efficiently a company produces its goods or services relative to its sales revenue. It is gross margin divided by sales revenue, reflecting the ability to manage production costs and generating profits from sales. Higher gross margin rate means the company is the leader of its industry, since they have higher-quality products or lower product costs compared to its competitors.

On the other hand, ROA evaluates how effectively to generate profits from a company's total assets. ROA is net income divided by total assets. Higher ROA means the company has better asset utilization and profitability.

We label the indicator by 3 different labels:

“Improve”: The indicator of the next year has increased by more than 1%.

“Unchanged”: The change of the indicator is between 1% and -1%.

“Decline”: The indicator of the next year is decreased by more than 1%.

C. Preprocessing

The preprocessing phase involves converting raw MD&A text data into a structured format for further analysis. We normalize the text into TF-IDF vectors for most classification models, and we transform the text into Multi-hot vectors for Bernoulli NB classifiers, TF vectors for Multinomial NB .

Take TF-IDF for example, we set the parameter “min_df=2”, which means that only words that appear in at least 2 documents will be included. Words that appear only in 1 document will be ignored. This can filter out rare words that are not meaningful for the analysis. We set the parameter “max_df=0.9”, which means that any word that appears in more than 90% of the documents will be excluded. This can remove very common words that might not be able to convey valuable information. In addition, we also remove stop words and convert texts into lowercase.

D. Models

We use several classification models, including Naive-Bayes, Bernoulli NB, Support Vector Machine(SVM), Rocchio, K-Nearest Neighbors (KNN). In addition, we also use multiple clustering models, such as K-means, Hierarchical Agglomerative Clustering(HAC). Moreover, we apply different dimensionality reduction methods such as Latent Dirichlet Allocation(LDA) and Singular Value Decomposition(SVD), and we combine them into different classification models. We also utilize recurrent neural networks(RNN) to analyze text in a sequential manner.

E. Precision

If the model predicts that a company will perform well while in the end the company gets worse, this will lead to huge loss in investment. Therefore, the cost of false positives is high, so we have to minimize false positives and maximize precision. As a result, we choose precision to measure the performance of our models.

F. Confusion Matrix and Merge Model

We also use the confusion matrix to understand what types of errors the models usually make. Based on the confusion matrix, we define a specific indicator called "non-fail rate", which will be introduced later. In the end, we combine some of the best-performing models to create a merge model. We incorporate dynamic weight adjustment and threshold setting to make the most confident predictions.

III. Research result

A. Gross Margin Precision

The precision values for predicting gross margin performance using various models are summarized in Table I. (G refers to growth companies, U refers to unchanging companies, D refers to declined companies)

As we can see, the precision of predicting growth companies in NB and KNN are the best three results since they all achieve precision values over 0.6. For unchanging companies, SVM stands out as the only model with precision higher than 0.6. Conversely, for declined-performing companies, CNB and SVM provide the highest precision. From a weighted average perspective, SVM is the most outperforming model.

Table 1. Gross Margin Performance allocation with 9 ways

	NB	Bernoulli NB	CNB	SVM	KNN	Rocchio	K-means	RNN	SVD+KNN
G	0.62	0.50	0.56	0.43	0.67	0.43	0.25	0.00	0.30
U	0.47	0.35	0.46	0.61	0.31	0.36	0.46	0.50	0.25
D	0.58	0.40	0.62	0.70	0.37	0.43	0.22	0.50	0.18
weighted avg	0.55	0.41	0.54	0.61	0.44	0.40	0.33	0.34	0.25

B. ROA Precision

The precision values for predicting ROA performance were similarly evaluated, which are summarized in Table II. In the case of ROA, SVM significantly outperforms other models, achieving perfect precision (1.00) for predicting growth companies. For unchanging companies, SVM, KNN, and Rocchio deliver notable results, with precision values exceeding 0.6. For declined-performing companies, KNN provides the highest precision. From a weighted average perspective, SVM and KNN are the most outperforming models.

There are almost no significant differences between NB models and SVM using Gross Margin or ROA as the mark of differing companies. But for KNN, Rocchio, RNN and KNN with SVD, ROA is much better than gross margin. For K-means, gross margin is better than ROA.

Table 2. ROA Performance allocation with 9 ways

	NB	Bernoulli NB	CNB	SVM	KNN	Rocchio	K-means	RNN	SVD+KNN
G	0.29	0.25	0.20	1.00	0.50	0.40	0.26	0.00	0.50
U	0.55	0.55	0.52	0.60	0.64	0.63	0.00	0.56	0.60
D	0.50	0.44	0.43	0.33	0.67	0.50	0.33	0.50	0.45
weighted avg	0.49	0.47	0.45	0.63	0.63	0.56	0.16	0.46	0.54

C. Confusion Matrix Analysis

The confusion matrixes of different models are presented below：(The row refers to the real categories of the companies, the column refers to the predicted categories of the companies.)

Table 3. Confusion Matrix of Gross Margin Using NB

	G	U	D
G	5	5	1
U	2	7	4
D	1	3	7

Table 4. Confusion Matrix of ROA Using NB

	G	U	D
G	2	3	0
U	4	12	3
D	1	7	3

Table 5. Confusion Matrix of Gross Margin Using Bernoulli NB

	G	U	D
G	4	5	2
U	3	6	4
D	1	6	4

Table 6. Confusion Matrix of ROA Using Bernoulli NB

	G	U	D
G	1	4	0
U	2	12	5
D	1	6	4

Table 7. Confusion Matrix of Gross Margin Using CNB

	G	U	D
G	5	5	1
U	3	6	4
D	1	2	8

Table 8. Confusion Matrix of ROA Using CNB

	G	U	D
G	1	3	1
U	4	12	3
D	0	8	3

Table 9. Confusion Matrix of Gross Margin Using SVM

	G	U	D
G	3	3	0
U	2	11	3
D	2	4	7

Table 10. Confusion Matrix of ROA Using SVM

	G	U	D
G	1	5	2
U	0	15	4
D	0	5	3

Table 11. Confusion Matrix of Gross Margin Using KNN

	G	U	D
G	2	5	4
U	1	4	8
D	0	4	7

Table 12. Confusion Matrix of ROA Using SVM

	G	U	D
G	2	3	0
U	1	16	2
D	1	6	4

Table13. Confusion Matrix of Gross Margin Using Rocchio

	G	U	D
G	3	5	3
U	3	5	5
D	1	4	6

Table15. Confusion Matrix of Gross Margin Using K-means

G	G	G	G
U	U	U	U
D	D	D	D

Table 16. Confusion Matrix of ROA Using K-means

	G	U	D
G	4	0	3
U	2	0	14
D	6	0	6

Table 14. Confusion Matrix of ROA Using Rocchio

	G	U	D
G	4	1	0
U	4	12	3
D	2	6	3

D. Non-failed rate analysis

They are presented in Table 17 and Table 18. These results highlight the models' effectiveness in classifying companies into performance categories with reasonable accuracy and show that at least they are less likely to make an opposite prediction.

Table 17. Non-failed rate analysis of Gross Margin

	NB	Bernoulli NB	CNB	SVM	KNN	Rocchio	K-means
non-failed rate(G)	0.88	0.88	0.89	0.71	1.00	0.86	0.67
non-failed rate(D)	0.92	0.80	0.92	1.00	0.79	0.79	0.25

Table 18. Non-failed rate analysis of ROA

	NB	Bernoulli NB	CNB	SVM	KNN	Rocchio	K-means
non-failed rate(G)	0.86	0.75	1.00	1.00	0.75	0.80	0.50
non-failed rate(D)	1.00	1.00	0.86	0.78	1.00	1.00	0.87

E. Model Combination

We developed a merged model that combines SVM and Naive Bayes to enhance overall performance metrics by dynamically adjusting weights and thresholds. The model weights were determined based on the respective accuracies of each model. Predictions were then made by incorporating these weights and their corresponding confidence levels. The merged model achieved a conservative precision and non-fail rates across both gross margin and ROA predictions. They are presented in Fig.1.

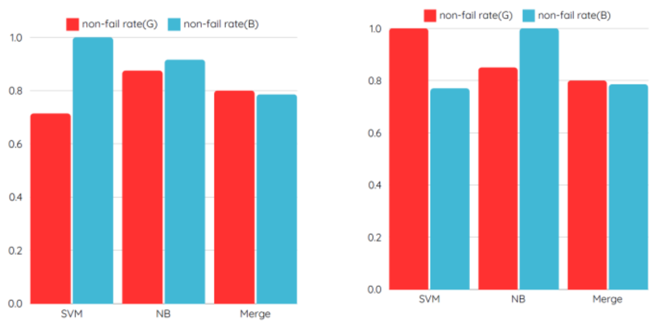

Fig. 1 Model Combination

IV. Conclusion

A. Improve Precision

In this study, we propose a novel approach to predict the future growth of companies using financial reports. Our models generate three possible outcomes: Growth (G), Stagnation (D), and Decline (B). However, when evaluating the performance of individual models based solely on the predicted class, the results were suboptimal. To address this, we introduced a confusion matrix to visualize the distribution of predictions. By considering stagnation as a successful prediction in the context of investment, we defined a new metric, non-fail rate, to assess model performance. Notably, some of our models achieved a non-fail rate of approximately 80%, demonstrating their potential to provide valuable insights for investment decisions.

B. Model ensemble techniques

Furthermore, we employed model ensemble techniques to enhance the precision and non-fail rate of our predictions. By combining the outputs of multiple well-performing models, we were able to improve the overall accuracy. We utilized confidence levels to select models with reliable predictions and dynamically adjusted their weights based on their performance. This approach ensured a more accurate and equitable ensemble prediction.

C. Future Work

This study aimed to develop a machine learning model capable of predicting corporate growth or decline based on textual content from financial reports, with the ultimate goal of providing valuable insights for investment decisions. The dataset comprised 342 financial reports, which, while providing a foundation for model development, highlights the need for a larger dataset to further enhance model performance. Our initial model classified companies into three categories: Growth (G), Stagnation (D), and Decline (B). Future research could explore a more granular classification system, such as G2, G1, D, B1, and B2, to provide more nuanced predictions. Additionally, we plan to incorporate traditional machine learning algorithms, such as Random Forest, to optimize model performance.

References

[1] Medya, S., Rasoolinejad, M., Yang, Y., & Uzzi, B. (2018, June). An exploratory study of stock price movements from earnings calls. In *Proceedings of Woodstock '18* (pp. 1–XX). Woodstock, NY.

[2] Feuerriegel, S., & Gordon, J. (2022). Long-term stock index forecasting based on text mining of regulatory disclosures. *ETH Zurich* and *University*

of Freiburg, Zurich, Switzerland, and Freiburg, Germany.

[3] Akash, B. S., & Dagli, C. (2022). Real-Time online stock forecasting utilizing integrated quantitative and qualitative analysis. *BITS Pilani Hyderabad Campus* and *Missouri University of Science and Technology*, 1–23.