Stock Volatility Forecasting and Word Features Selection — Based On 10-K Corpus

WANG Zeyu LIU Chang TAN Huangao

LI Borui ZHU Xinqi GAO Yifeng

March 23, 2025

Abstract

As a key indicator of the market risk, the stock volatility is an important evidence to support the investment portfolio. The traditional time series models such as GARCH models only rely on the volatility, but ignore the unstructured text data like companies' 10-K reports, which is also shown to be helpful in forecasting. However, due to the complexity of natural language and the evolving semantic patterns, it is still challenging to extract operable signals from the narration of natural language. Focusing on exploring the relation between the 10-K report and stock volatility, we apply Lasso regression, Ridge regression, and the decision tree on the companies' 10-K reports together with historical volatility, to predict the future volatility. The result shows that it can perform much better than the Naive model. Meanwhile, based on the word features given by the models, we also explore the difference of word features between the years and models and found it changes due to some big events.

Keywords: Lasso; Ridge; decision tree; stock volatility; natural language processing.

1 Introduction

1.1 Motivation and Goal

As a key property to measure market risk, stock volatility includes uncertainty and different expectations of investor for stock investment [1], which is important for portfolio [2]. In the internet era, the rapidly growing unstructured text data, especially in the aspect of company disclosure, enables the model to rely more on unstructured text data like the reports or acts [3] rather than structured data like historical prices and volatility, which provides great potential for improve volatility prediction [4] . Even though the previous research has preliminarily shown the link between text sentiment and short-term market response [5], it is still challenging to extract operable signals from the narration of natural language due to the complexity of natural language and the evolving semantic patterns in the annual report.

Based on the above motivation, we hope to predict the logarithmic volatility of stock returns in the next year by studying the management discussion and analysis (MD&A) part of the form 10-K report [6] [7], as well as identify and evaluate the importance of a single token in volatility prediction. The machine learning model including Lasso/Ridge regression and decision tree, combining with the NLP technology [8] will be used to establish the framework for text driven volatility prediction, reveal language patterns, and provide effective information for investment strategies, company disclosures and regulatory practices.

1.2 Datasets

The 10-K Corpus [9] is a widely used dataset including the 10-K reports from thousands of companies, with the stock return volatility measurements in the one year period before and after each report. Its subset is used in our work, including the log volatility (e.g., 1996.logvol.-12.txt), as well as the tokenized MD&A sections (e.g., 1996.tok.tgz), which provides us with the detailed

discussion about the operations of the company and some forward looking statements [10]. As for the other data, such as the meta data about the companies, are regarded irrelevant to our goals.

2 Methods

2.1 Text preprocessing

We use the Natural Language Toolkit (NLTK) [11] as lemmatizer, stemmer and stop token filter with the idea that: (1) Some words including articles, prepositions and conjunctions are regarded as insignificant in text analysis but appears in every report; (2) Words appear in several inflected forms but with same meaning should be considered as the same, and should be transformed to their stem, so that they can be clustered together. Through these text preprocessing methods, we can greatly reduce the text volume, eliminate the noise, and improve the efficiency of the following steps.

2.2 Term Frequency–Inverse Document Frequency (TF-IDF)

The term frequency–inverse document frequency (TF-IDF) [12] [13] measures the importance of a word to a document in corpus via

$$\text{(Term frequency)} \qquad \text{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

$$\text{(Inverse document frequency)} \quad \text{idf}(t,d) = \log_2 \left(\frac{N}{|\{d : d \in D, t \in d\}|} \right),$$

where $f_{t,d}$ is the raw count of a term t in a document d, and N is the total number of documents. With the TF-IDF, each company's 10-K report is represented by a vector that highlights discriminative keywords.

2.3 Nonnegative Matrix Factorization (NMF)

In order to further reduce the dimension of the input data, we use the nonnegative matrix factorization (NMF) [14], with which we can decompose the TF-IDF matrix $X \in \mathbb{R}^{n \times m}$ into two nonnegative smaller matrices $W \in \mathbb{R}^{n \times d}$ and $H \in \mathbb{R}^{d \times m}$, where $X \approx WH$ and d is the given number of dimension.

The previous research [15] have shown that this method is particularly suitable for text data due to non-negativity and is able to learn the semantic features of text, which allows us to extract interpretable latent topics that capture the underlying themes in the MD&A sections. Besides, unlike other dimensionality reduction techniques like PCA, NMF provides sparse and parts-based representations, making it easier to identify meaningful topics and understand the contribution of individual words.

2.4 Naive Model (Baseline)

The previous research [16] has shown that the naive model is better than many other complex models, especially for high-volatility time series. Thus we choose the naive model (or Persistence model) as our baseline, which uses the current (or same period) value as the forecast. Since it only requires the doesn't involve the word features or any assumptions.

2.5 Lasso/Ridge Regression & Decision Tree

The models we used are Lasso regression, Ridge regression and Decision Tree, which are all widely known. These models (also their extensions) are powerful in forecast and can have better performance compared with other models like heterogeneous autoregressive (HAR) models [17] [18] [19]. In addition, they are well-known inherently interpretable, which can help us extract whether a word implies the increasing or decreasing of volatility.

2.6 Model Evaluation

The two widely used measurement MAE and MSE will be used to evaluate the results. Given the prediction $\hat{x}$ and ground truth x, they are computed as

$$\text{MAE}(x, \hat{x}) = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|, \quad \text{MSE}(x, \hat{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2.$$

Another problem we are interested in is the words features. We first compute the weight of each words via the pseudo inverse of H with $H^+\beta = (V^T\Lambda^+U)\beta$, where $\beta \in \mathbb{R}^d$ is the model parameter and $H \in \mathbb{R}^{d\times m}$ is components by NMF. Here the pseudo inverse H^+ is calculate from the SVD that $H = U^T\Lambda V$, and Λ^+ is the diagonal matrix consisting of the reciprocals of H's singular values. In order to compare the features cross models and years, we selected the top 1000 words (about $1\% \sim 5\%$ of total) for each model and check whether they are similar between different methods and years.

2.7 Overview

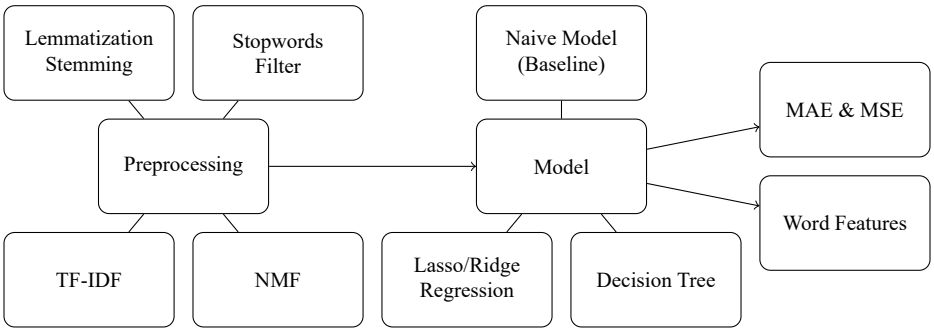

Figure 1: Overview flowchart for our method.

The Figure 1 is an overview of our method. The number of the components by NMF is set to 20 balancing the memory, time and the accuracy. Consider the number that coefficients close to zero and the model error, the penalty is 0.00005 for Lasso and 1.0 for Ridge, meanwhile, the depth

of the tree is set to 5.

Then we test the model in two approaches. First we choose the data from one or several years, and split them into training dataset (80%) and test dataset (20%). In this case, the model is trained and tested on the data from same years, thus to avoid the flukes caused by splitting, we repeat the experiment for 400 times, with the central limit theorem and sample variance, we can control the error under 0.003 (with the probability over 95%), which is accurate enough.

Another test is to train the model via the data from one or several years and forecast the following year. In this case, we use all the data from previous years as the training set and the whole following year's data as the testing set. This approach will be run once to show the model's ability to deal with the new text data.

3 Result

3.1 Forecasting error

3.1.1 Error with 1-year's training data

We first train the data with 1-year text data and volatility. The Figure 2 shows the error of the model predicting the volatility in the same year, while the Figure 3 shows the error of the prediction with the text and volatility of the next year.

From the result, we can infer that the Lasso and Ridge regression can get almost the same error while the error of the decision tree is much higher. And all the models perform better when predicting the volatility of the same year rather than the following year. Furthermore, the prediction of the following year's volatility, especially the result of Lasso and Ridge regression, achieves much more accurate after 2002 with the enforcement of Sarbanes-Oxley Act, which implies that the this act might make the report more informative. Another noticeable difference is the error of predicting the following years' volatility gets higher in 2000, which may be caused by the dot-com bubble.

Figure 2: MAE and MSE for predicting the 1-year's data with the model trained by the same 1-year's data.

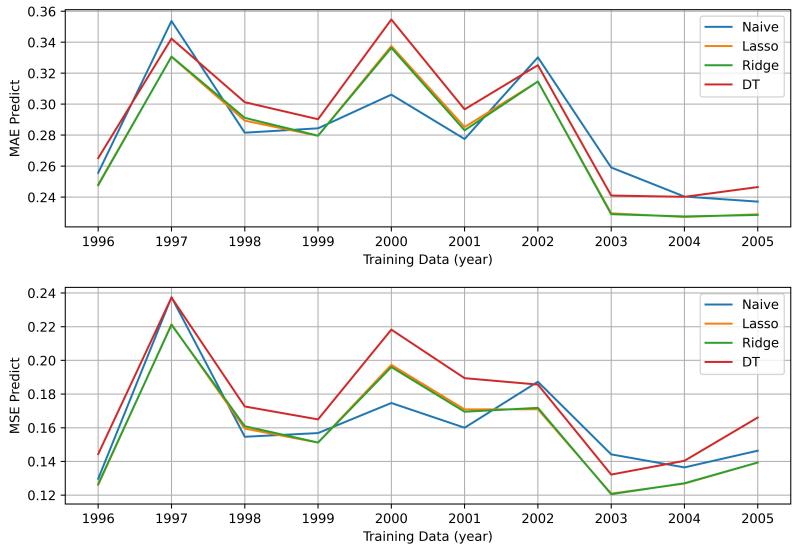

Figure 3: MAE and MSE for predicting the following 1-year's data with the model trained by the previous 1-year's data.

3.1.2 Error with 3-years' training data

It is widely known that more training data can lead to a better method performance, with the assumption that the data are from the same distribution. In this case, we presume that the information or the distribution of the words are similar across the nearby years. This assumption is reasonable, otherwise we can't expect the model to do the prediction. But it still needs to be certified.

Therefore we train the model with the data from nearby 3 year, and test the model on the same 3 year's data as well as the following year's data. The Figure 4 shows the error of the model predicting the volatility in the same 3 years, while the Figure 5 shows the error of the prediction with the text and volatility of the following year.

Similar to the result of 1-year's training data, the result still shows the influence of the dot-com bubble and the Sarbanes-Oxley Act. Meanwhile, the errors of three models become more closer and more prediction are better than or equivalent to the Naive model, which implies the higher stability and higher accuracy than the model based on the 1-year's data.

3.2 Word Features

Another problem we concern about is the words feature, especially the difference and similarly across years or models. Thus we compare the words features given by different models with training data from different years. The Table 1 compares the words features across years, where the column "Increasing Common" and "Decreasing Common" shows the rate of the nearby two years' common features related to the increasing or decreasing volatility, and the column "Different" shows the rate of the features that are corresponding to the different trend, i.e. the rate of the features that change the meanings in two years.

This table shows that in the most of cases, the word features are similar between the nearby years, while there are some abnormal values, mostly around 2000 and 2003. The the abnormal values are corresponding to the year of the dot-com bubble and the Sarbanes-Oxley Act, which

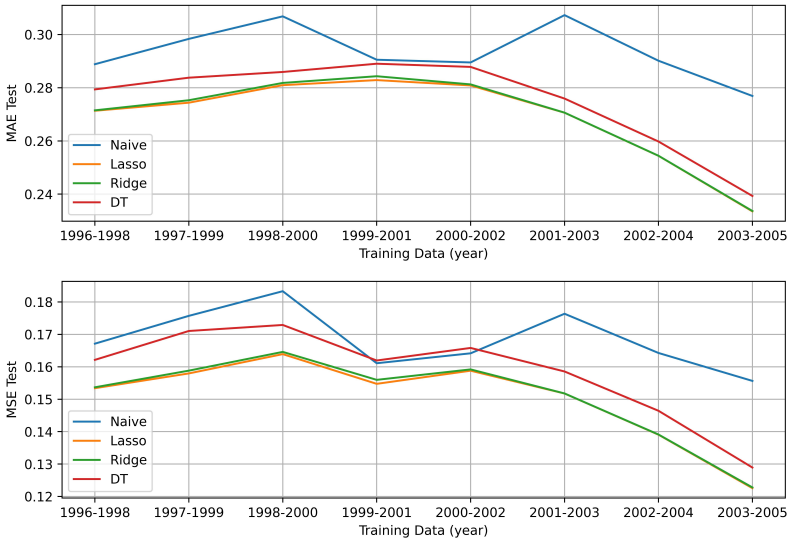

Figure 4: MAE and MSE for predicting the 3-years' data with the model trained by the same 3-years' data.

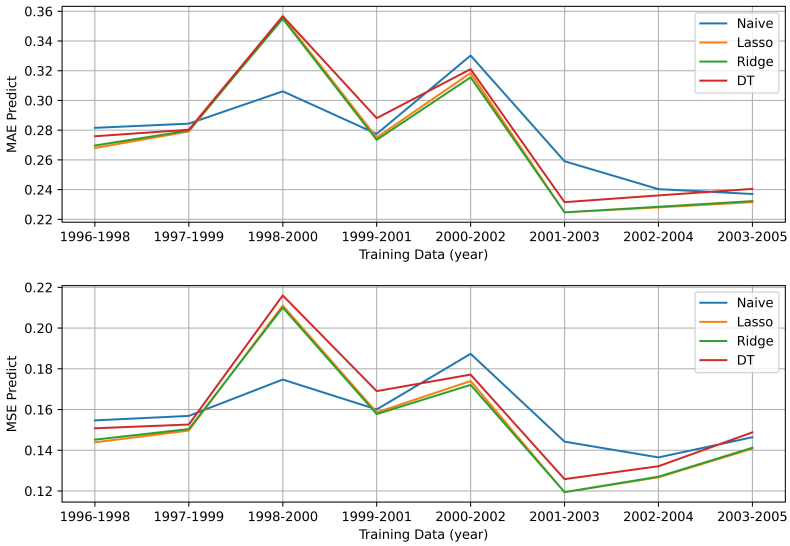

Figure 5: MAE and MSE for predicting the following 1-year's data with the model trained by the previous 3-years' data.

Year	Method	Increasing Common	Decreasing Common	Different
1996-1997	Lasso	53.9%	60.6%	1.9%
	Ridge	48.3%	55.1%	2.9%
	Decision Tree	62.7%	15.8%	3.6%
1997-1998	Lasso	35.1%	39.2%	6.4%
	Ridge	38.1%	36.9%	7.55%
	Decision Tree	46.2%	29.4%	6.75%
1998-1999	Lasso	23.6%	31.7%	5.6%
	Ridge	20.2%	32.1%	8.35%
	Decision Tree	37.1%	18.8%	8.4%
1999-2000	Lasso	3.8%	27.8%	17.1%
	Ridge	9.9%	38.6%	6.55%
	Decision Tree	29.3%	33.2%	4.4%
2000-2001	Lasso	5.1%	24.9%	1.2%
	Ridge	13.2%	18.9%	3.25%
	Decision Tree	23.9%	32.7%	6.2%
2001-2002	Lasso	38.0%	30.4%	2.7%
	Ridge	42.2%	35.7%	4.15%
	Decision Tree	23.7%	3.6%	10.9%
2002-2003	Lasso	31.7%	31.0%	12.15%
	Ridge	35.8%	36.4%	9.2%
	Decision Tree	45.7%	5.8%	4.3%
2003-2004	Lasso	26.2%	58.3%	3.15%
	Ridge	27.3%	44.6%	1.25%
	Decision Tree	34.7%	2.0%	11.85%
2004-2005	Lasso	42.2%	38.3%	6.4%
	Ridge	45.5%	43.6%	9.3%
	Decision Tree	56.7%	7.3%	2.1%

Table 1: The rate of common and different word features across different years. The column "Increasing Common" and "Decreasing Common" shows the rate of the nearby two years' common features related to the increasing or decreasing volatility, the column "Different" shows the rate of the features that are corresponding to the different trend, i.e. the rate of the features that changes the meanings in two years. Values in red are outliers from the whole, and blue means the large gap from the previous and following year.

confirm the conclusion that the big events will affect the information or the distribution of the words, and it also shows that the firm are more likely to follow the act since 2003.

The Table 2 compares the word features given by different models. Similar to the previous, the rates of the common features and different features shows that the Lasso and Ridge regression shares most of the features, while the decision tree gives the totally different features.

We try to figure out the reason behind this phenomena by analyse the coefficients given by the models. The parameter shows that the decision tree puts a higher importance on the historical volatility, while the Lasso and Ridge regression gives a much more averagely coefficients. For example, with the training data from 1996, the decision tree gives all the word importance less than 0.1, while those given by the Lasso and Ridge regression can be up to 0.5. This might be caused by many reasons, but we believe that the most important one is that the decision tree is in fact non linear, thus when it tries to fit the data, it will obviously concern more on the historical volatility as it will be much more significantly different compare with the word features. One possible way to reduce this is to predict the difference of the volatility, or simply do the classification on the trend, which may improve the preformance in further research.

4 Conclusion

We applied the Lasso regression, Ridge regression, and decision tree to the companies' 10-K report, as well as the historical volatility, to forecast the volatility. Similar to the previous research [9], our result also shows that the 10-K report, especially the report after the Sarbanes–Oxley Act, can help to predict the value better than the naive model with the similar word features across models and years, and more training data can lead to a better model performance. Furthermore, big events like the dot-com bubble and the Sarbanes-Oxley Act can make a significant effect on the information or the distribution of the text.

Year	Compared Methods	Increasing Common	Decreasing Common	Different
1996	Lasso & Ridge	75.2%	71.6%	0.15%
	Lasso & Decision Tree	23.2%	7.7%	27.0%
	Ridge & Decision Tree	36.1%	25.1%	17.35%
1997	Lasso & Ridge	82.0%	87.2%	0.15%
	Lasso & Decision Tree	15.4%	7.3%	25.85%
	Ridge & Decision Tree	16.6%	12.2%	25.6%
1998	Lasso & Ridge	84.4%	85.5%	0.2%
	Lasso & Decision Tree	14.1%	15.1%	15.4%
	Ridge & Decision Tree	12.2%	15.6%	18.6%
1999	Lasso & Ridge	67.9%	83.9%	0.35%
	Lasso & Decision Tree	17.8%	23.8%	19.8%
	Ridge & Decision Tree	11.5%	23.1%	24.0%
2000	Lasso & Ridge	68.1%	55.3%	1.85%
	Lasso & Decision Tree	19.0%	25.1%	17.6%
	Ridge & Decision Tree	24.1%	9.5%	26.1%
2001	Lasso & Ridge	91.7%	92.9%	0.1%
	Lasso & Decision Tree	5.5%	0.9%	31.55%
	Ridge & Decision Tree	6.8%	1.6%	30.55%
2002	Lasso & Ridge	88.6%	79.1%	0.1%
	Lasso & Decision Tree	37.7%	49.8%	9.75%
	Ridge & Decision Tree	33.8%	39.2%	14.35%
2003	Lasso & Ridge	93.6%	89.5%	0.15%
	Lasso & Decision Tree	13.4%	10.4%	27.05%
	Ridge & Decision Tree	15.2%	13.9%	25.45%
2004	Lasso & Ridge	75.2%	73.4%	0.2%
	Lasso & Decision Tree	23.5%	24.2%	21.25%
	Ridge & Decision Tree	30.4%	20.4%	20.85%
2005	Lasso & Ridge	95.6%	90.8%	0.4%
	Lasso & Decision Tree	9.9%	2.5%	27.5%
	Ridge & Decision Tree	10.2%	5.2%	28.7%

Table 2: The rate of common and different word features across different models. The column "Increasing Common" and "Decreasing Common" shows the rate of the nearby two years' common features related to the increasing or decreasing volatility, the column "Different" shows the rate of the features that are corresponding to the different trend, i.e. the rate of the features that change the meanings in two years. Values in red are outliers from the whole, and blue means the large gap from the previous and following year.

References

[1] Le Zhao, Vinh Huy Nguyen, and Chen Li. The volatility-liquidity dynamics of single-stock etfs. *Finance Research Letters*, 69(PB), 2024.

[2] Van-Dai Ta, Chuan-Ming Liu, and Direselign Addis Tadesse. Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading. *Applied Sciences*, 10(2):437, 2020.

[3] Hiridik Rajendran, Parthajit Kayal, and Moinak Maiti. Is the u.s. energy independence and security act of 2022 associated with stock market volatility? *Utilities Policy*, 90:101813, 2024.

[4] Yong Ma, Shuaibing Li, and Mingtao Zhou. Twitter-Based Market Uncertainty and Global Stock Volatility Predictability. *The North American Journal of Economics and Finance*, 72:102256, 2024.

[5] Jinfang Li. The sentiment pricing dynamics with short-term and long-term learning. *The North American Journal of Economics and Finance*, 63:101812, 2022.

[6] Raffaele Mattera and Philipp Otto. Network log-arch models for forecasting stock market volatility. *International Journal of Forecasting*, 40(4):1539–1555, 2024.

[7] Feng Ma, Jiqian Wang, MIM Wahab, and Yuanhui Ma. Stock market volatility predictability in a data-rich world: A new insight. *International Journal of Forecasting*, 39(4):1804–1819, 2023.

[8] Dani Yogatama. *Sparse models of natural language text*. PhD thesis, Ph. D. thesis, Carnegie Mellon University, 2015.

[9] Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. Predicting risk from financial reports with regression. In *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, 2009.

[10] Stephen V Brown, Lisa A Hinson, and Jennifer Wu Tucker. Financial statement adequacy and firms' md&a disclosures. *Contemporary Accounting Research*, 41(1):126–162, 2024.

[11] Bird, Steven, Edward Loper, and Ewan Klein. *Natural Language Processing with Python.* O'Reilly Media Inc., 2009.

[12] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[13] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

[14] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

[15] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

[16] Nico Beck, Jonas Dovern, and Stefanie Vogl. Mind the naive forecast! a rigorous evaluation of forecasting models for time series with low predictability. *Applied Intelligence*, 55(6):395, 2025.

[17] Chao Liang, Yongan Xu, Zhonglu Chen, and Xiafei Li. Forecasting china's stock market volatility with shrinkage method: Can adaptive lasso select stronger predictors from numerous predictors? *International Journal of Finance & Economics*, 28(4):3689–3699, 2023.

[18] Xiafei Li, Chao Liang, and Feng Ma. Forecasting stock market volatility with a large number of predictors: New evidence from the ms-midas-lasso model. *Annals of Operations Research*, pages 1–40, 2022.

[19] Kim Christensen, Mathias Siggaard, and Bezirgen Veliyev. A machine learning approach to volatility forecasting. *Journal of Financial Econometrics*, 21(5):1680–1727, 2023.

A Author Contributions

- WANG Zeyu: Write and test the code for processing, Lasso and Ridge regression; Write the Result and Conclusion sections of the report.
- LIU Chang: Write and test the code for Decision tree; Write the Method section of the report.
- TAN Huangao: Write and test the code for text processing; Write the Method section of the report.
- LI Borui: Test the code; Write the Introduction section of the report.
- ZHU Xinqi: Write the Introduction section of the report; Make the slide; Do the presentation.
- GAO Yifeng: Write the Find section of the report; Make the slide; Do the presentation.