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 22, 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: keywords

1 Introduction

The effective transmission and interpretation of qualitative information in financial markets play a key role in shaping investor behavior and dynamic asset pricing. As a key indicator to measure market risk, stock volatility includes uncertainty and different expectations of investor for stock investment [1] . As the traditional quantitative models mainly rely on structured financial data (such as historical prices and accounting ratios), unstructured text data has shown an explosive growth trend in the Internet era, especially in the aspect of company disclosure, which provides untapped potential for enhancing volatility prediction [2] . Although it has information value, due to the complexity of natural language and the evolving semantic patterns in the annual report, it is still challenging to extract operable signals from the narration of natural language. Previous studies have established a preliminary link between text sentiment and short-term market response [3] , but the dynamic interaction between lexical features and long-term fluctuations, as well as the temporal stability of these features, have not been fully explored. This gap not only limits the improvement of prediction model, but also hinders the optimization of corporate disclosure practice and regulatory supervision.

1.1 Motivation and Goal

The motivation for this study is driven by four interconnected objectives that bridge practical financial applications, corporate governance, and academic innovation. First, we seek to decode market information transmission mechanisms by analyzing the impact of unstructured textual data in 10-K reports — particularly the Management's Discussion and Analysis (MD&A) section [4] — on stock volatility. This investigation validates whether financial markets rapidly and accurately internalize qualitative disclosures, thereby shedding light on how investors interpret textual information to inform trading decisions. Second, our work aims to construct quantitative investment strategies by identifying key lexical features within 10-K filings. By developing data-driven mod-

els to predict short-term stock volatility and flagging risk-indicative terms (e.g., negative sentiment words), we provide tools for portfolio optimization and preemptive risk mitigation [5] . Third, this research addresses corporate governance and disclosure quality enhancement [6] . By pinpointing textual patterns that disproportionately influence market reactions — such as ambiguous or exaggerated language — we offer actionable insights for firms to refine their communication strategies. Concurrently, regulators can leverage these findings to detect deliberate obfuscation of risks in corporate disclosures, fostering greater transparency. Finally, this study contributes to academic research by pioneering methodologies at the intersection of finance and computational linguistics. By advancing text-based volatility forecasting frameworks, we enrich interdisciplinary dialogue and establish novel paradigms for analyzing financial narratives. Collectively, these objectives underscore the transformative potential of integrating textual analytics into financial decision-making, risk management, and regulatory oversight.

Based on the motivation of the above experiment, this study mainly pursues the following research objectives. First of all, we hope to predict the logarithmic volatility of stock returns in the next year by studying the MD&A part of the form 10-K report [7] [8] . We will use the machine learning model including Lasso/Ridge regression and decision tree, and combine NLP technology [9] to extract semantic features from unstructured text. The second goal is to identify and evaluate the importance of a single token in predicting volatility, focusing on its time stability. Through pseudo-inverse decomposition of NMF components [10] and feature selection metrics, we separated the first 1000 words that were most closely related to volatility changes. This allows us to check whether the forecast terms (e.g., risk related phrases) are consistent in different years or change with regulatory changes.

These objectives aim to establish an effective framework for text driven volatility reduction, reveal operable language patterns, and provide effective information for investment strategies, company disclosures and regulatory practices.

1.2 Datasets

To achieve these objectives, we leverage a curated subset of the 10-K Corpus [11] , spanning annual filings from 1996 to 2005.In the original data set, the data of these two parts have better value for research. First, the tokenized MD&A sections(e.g., 1996.tok.tgz), which is the most directly related component used for prediction in Form 10-K.It provides the insights of management on the financial condition of company, changes in financial condition and overall operation. This section contains forward-looking statements, risk disclosures and strategic insights, which are not available in other parts of the report. Therefore, this section provides a rich source of information for predicting future stock volatility. Feature extraction can help us understand the transmission mechanism of market information. The second part is corresponding annualized log volatility metrics derived from stock returns, aligned temporally with the 10-K reporting cycle(e.g., 1996.logvol.-12.txt).

2 Methods

2.1 Text preprocessing

We use the Natural Language Toolkit (NLTK) [12] as lemmatizer, stemmer and stop token filter with the idea that: (1) Some words are regarded as insignificant in text analysis, e.g. a, an, the, above, across; (2) Words appear in several inflected forms but with same meaning should be considered as the same, e.g. the words achieves, achieved, achievement, achievements are the same and we transform them to the root achiev. This method can greatly reduce our data processing volume and help to reduce the memory and time usage.

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

The term frequency–inverse focument frequency (TF-IDF) [13] [14] 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. After computing the TF-IDF, each company is represented by a vector where each element corresponds to a token's TF-IDF weight, highlighting 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) [15] , with which we can decomposes 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 [16] 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 [17] 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. These models (also their extensions) are powerful in forecast and can have better performance compare with other models like heterogeneous autoregressive (HAR) models [18] [19] [20]. 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 where they are computed as

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

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.

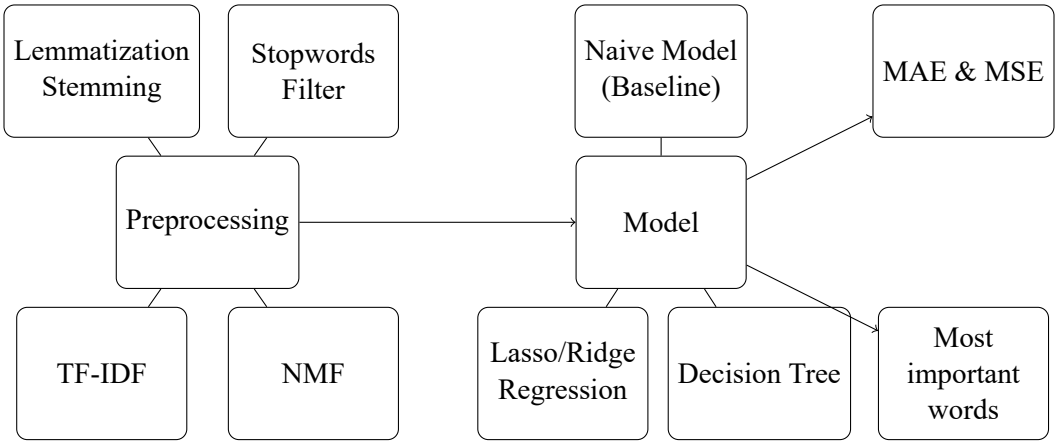

Figure 1: Overview flowchart for our method.

2.7 Overview

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 for the Lasso is 0.00005 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 cause 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 enough accurate.

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 Model trained by 1-year 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 easily infer that the Lasso regression 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 data of the same year rather than the following year. Furthermore, the prediction of the following year, 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 different is that the error of predicting the following year gets worse in 2000, which may cause by the dot-com bubblt.

3.1.2 Model trained by 1-year training data

It is widely known that more training data can lead to a better method performance, with the assumption that all 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 next 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.

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

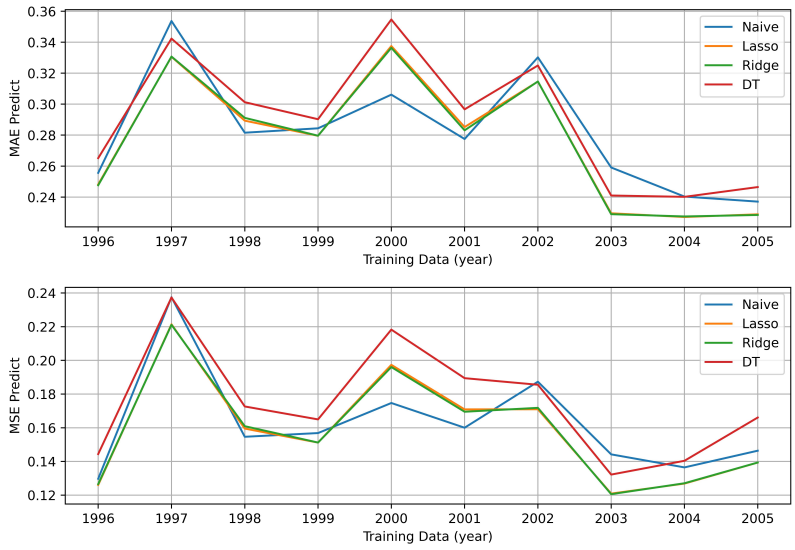

Figure 3: MAE and MSE for predicting the following 1-year's data with the model trained by the previous 1-year's data.

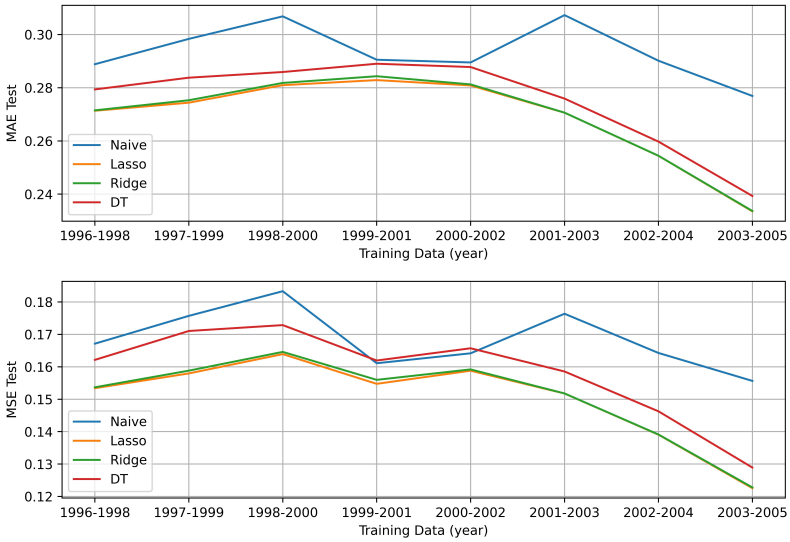

Figure 4: MAE and MSE for predicting the 3-year's data with the model trained by the same 3-year's data.

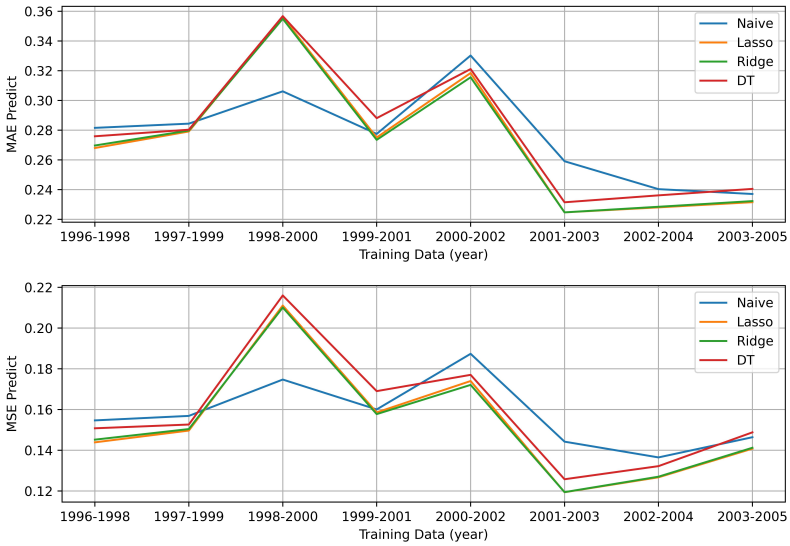

Figure 5: MAE and MSE for predicting the following 1-year's data with the model trained by the previous 3-year's data.

Similar to the 1-year 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 Wrod Features

Another problem we concern about is the words feature, especially whether they are different cross 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 common features related to the increasing or decreasing volatility from nearby two years, the column "Different" shows the rate of the features that are corresponding to the different trend in the two years, i.e. the rate of the features that changes the meanings.

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

The Table 2 compares the word features given by different models. Similar to the previous, we compute the rate of the common features and different features. It is easy to see 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 decisionn tree puts a higher importance on the historical volatility, while the Lasso and Ridge regression gives a much more averagely coefficients. For

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

example, with the training data from 1996, the decision tree will give all the importance of word features 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 can be used 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 [21], 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, 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 common features related to the increasing or decreasing volatility from nearby two years, the column "Different" shows the rate of the features that are corresponding to the different trend in the two years, i.e. the rate of the features that changes the meanings.

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] 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.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, and Wing-Kin Ma. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Process. Mag.*, 36(2):59–80, 2019.

[11] 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.

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

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

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

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

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

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

A Author Contributions