Explore the Influence of Words on Stock Volatility and the Most Important Word Features — Based On 10-K report

WANG Zeyu
LIU Chang
TAN Huangao
LI Borui
ZHU Xinqi
GAO Yifeng

March 16, 2025

Contents

Introduction - Goals

Our project aims to:

- Predict the log volatility of the following year via the MD&A section of the Form 10-K;
- Find out the importance of each token and check whether they are similar over the years.

Introduction - Motivation

These two goals are meaningful because:

- The MD&A section is most related to the forecasting in Form 10-K;
- A better predicting model can lead to a better portfolio;
- Such models show the market information transmission mechanisms.

Introduction - Dataset

We use a portion of the data in 10-K Corpus[1], including:

- The tokenized MD&A sections;
  - (e.g. 1996.tok.tgz)
- The log volatility in the related year.
  - (e.g. 1996.logvol.-12.txt, 1996.logvol.+12.txt)

[1] Shimon Kogan et al. (2009). "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*, pp. 272–280.

Methods - Lemmatization, Stemming and Stop token filter

We use the Natural Language Toolkit (NLTK)[2] for lemmatizer, stemmer and stop token filter.

- Some words are regarded as insignificant in text analysis;
  - (e.g. a, an, the, above, across)
- Words appear in several inflected forms but with same meaning should be considered as the same;
  - (e.g. achieves, achieved, achievement, achievement $\rightarrow$ achiev)

[2] Bird et al. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

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

The term frequency–inverse focument frequency (TF-IDF)[3][4] measures the importance of a word to a document in corpus via:

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

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

where $f_{t,d}$ is the raw count of a term t in a document d.

[3] Karen Sparck Jones (1972). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* 28.1, pp. 11–21.

[4] Stephen Robertson (2004). "Understanding inverse document frequency: on theoretical arguments for IDF". In: *Journal of documentation* 60.5, pp. 503–520.

Methods - Nonnegative Matrix Factorization (NMF)

Given a matrix $A \in \mathbf{n} \times \mathbf{m}$, the nonnegative matrix factorization (NMF)[5] aims to find two nonnegative matrixs $W \in \mathbf{n} \times \mathbf{d}$, $H \in \mathbf{d} \times \mathbf{m}$, where d is the given number of dimension, such that

$$V \approx WH.$$

The previous research[6] have shown that this method is able to learn the semantic features of text.

[5] Daniel Lee and H Sebastian Seung (2000). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems* 13.

[6] Daniel D Lee and H Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". In: *nature* 401.6755, pp. 788–791.

Methods - Naive Model (Baseline)

- The naive model (or Persistence model) uses the current (or same period) value as the forecast;
- The previous research[7] has shown that the naive model is better than many other complex models, especially for high-volatility time series;
- It can't show how the features affect the forecast.

[7] Nico Beck, Jonas Dovern, and Stefanie Vogl (2025). "Mind the naive forecast! a rigorous evaluation of forecasting models for time series with low predictability". In: *Applied Intelligence* 55.6, p. 395.

Methods - Lasso/Ridge Regression & 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[8][9][10];
- These models are inherently interpretable, which can help us extract whether a word implies the increasing or decreasing of volatility.

[8] Chao Liang et al. (2023). "Forecasting China's stock market volatility with shrinkage method: Can Adaptive Lasso select stronger predictors from numerous predictors?" In: *International Journal of Finance & Economics* 28.4, pp. 3689–3699.

[9] Xiafei Li, Chao Liang, and Feng Ma (2022). "Forecasting stock market volatility with a large number of predictors: New evidence from the MS-MIDAS-LASSO model". In: *Annals of Operations Research*, pp. 1–40.

[10] Kim Christensen, Mathias Siggaard, and Bezirgen Veliyev (2023). "A machine learning approach to volatility forecasting". In: *Journal of Financial Econometrics* 21.5, pp. 1680–1727.

Methods - Model Evaluation

We use mean absolute error and mean squared error to evaluate the results:

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

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

Methods - Model Evaluation

Given the model parameter $\beta \in \mathbf{R}^d$ and NMF components $H \in \mathbf{R}^{d \times m} = (h_1, \ldots, h_m)$, we compute

$$\beta^T h_i$$

for the i-th word as the overall weight, and choose the words with maximal/minimal weight for the features of increasing/decreasing.
We selected the top 100 words and check whether they are similar between different methods and years.

Methods - Overview Flowchart

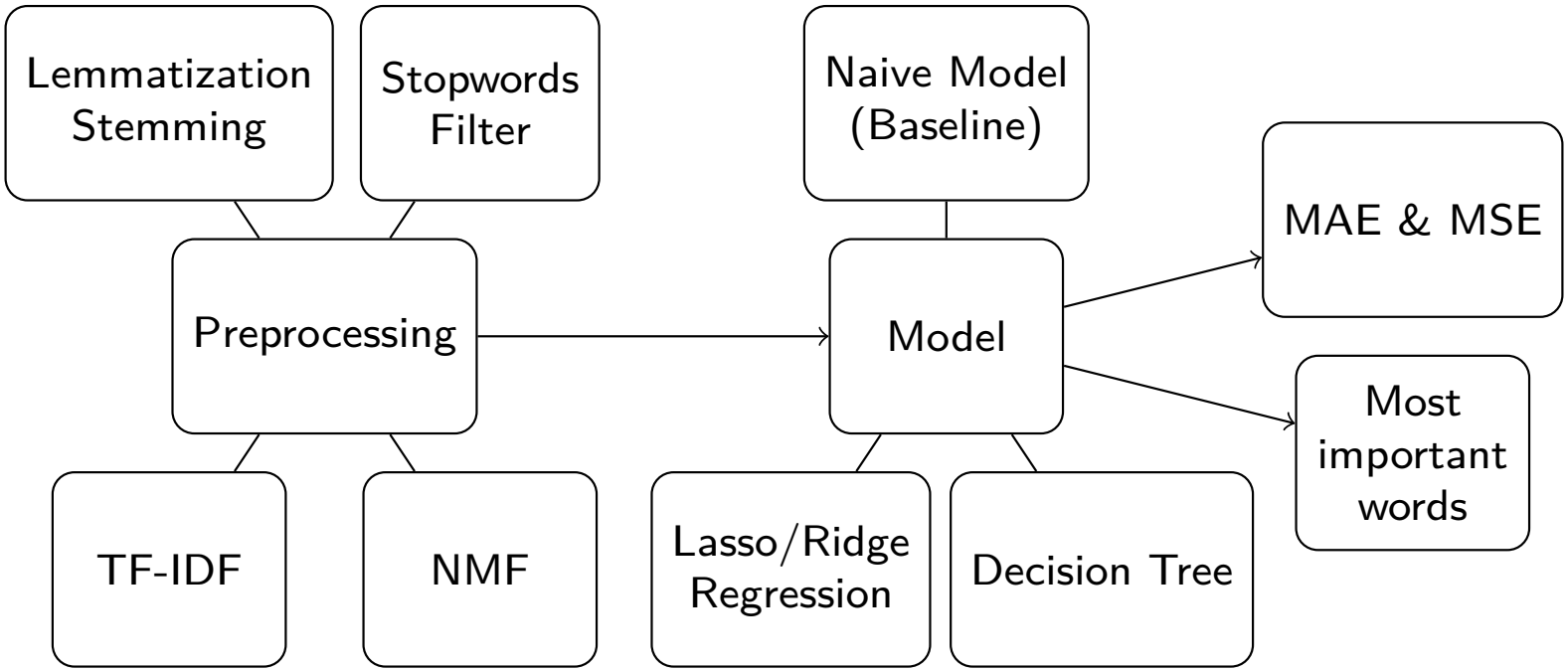

Figure: Flowchart for our methods.

We test the model in two approaches:

- Split the data from some years into training dataset (80%) and test dataset (20%), and repeat the experiment to avoid flukes;
- Train the model via the data from some years and forecast the following year.

Result - Error

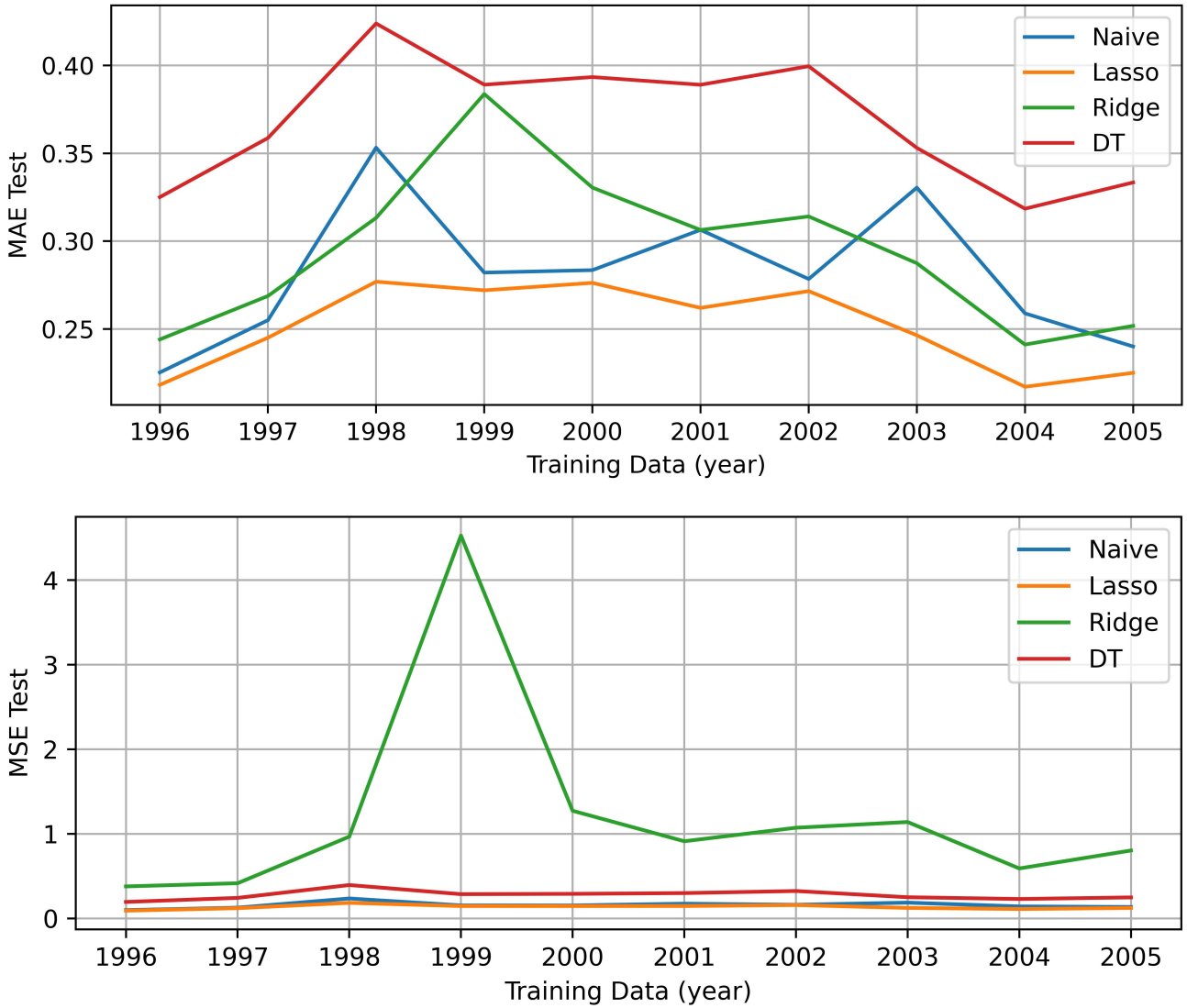

Figure: MAE and MSE for testing 1-year training set.

Result - Error

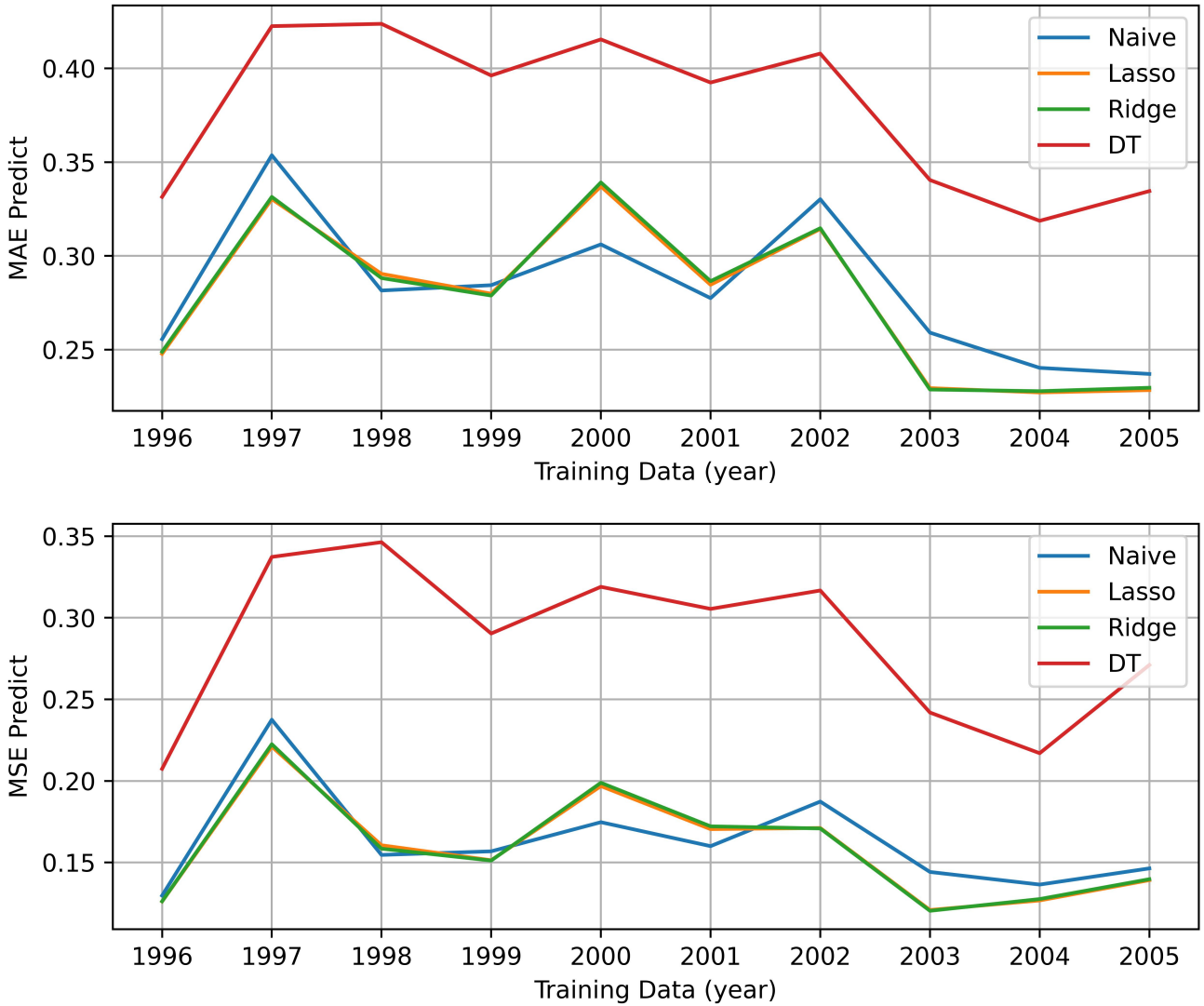

Figure: MAE and MSE for predicting 1-year training set.

Result - Error

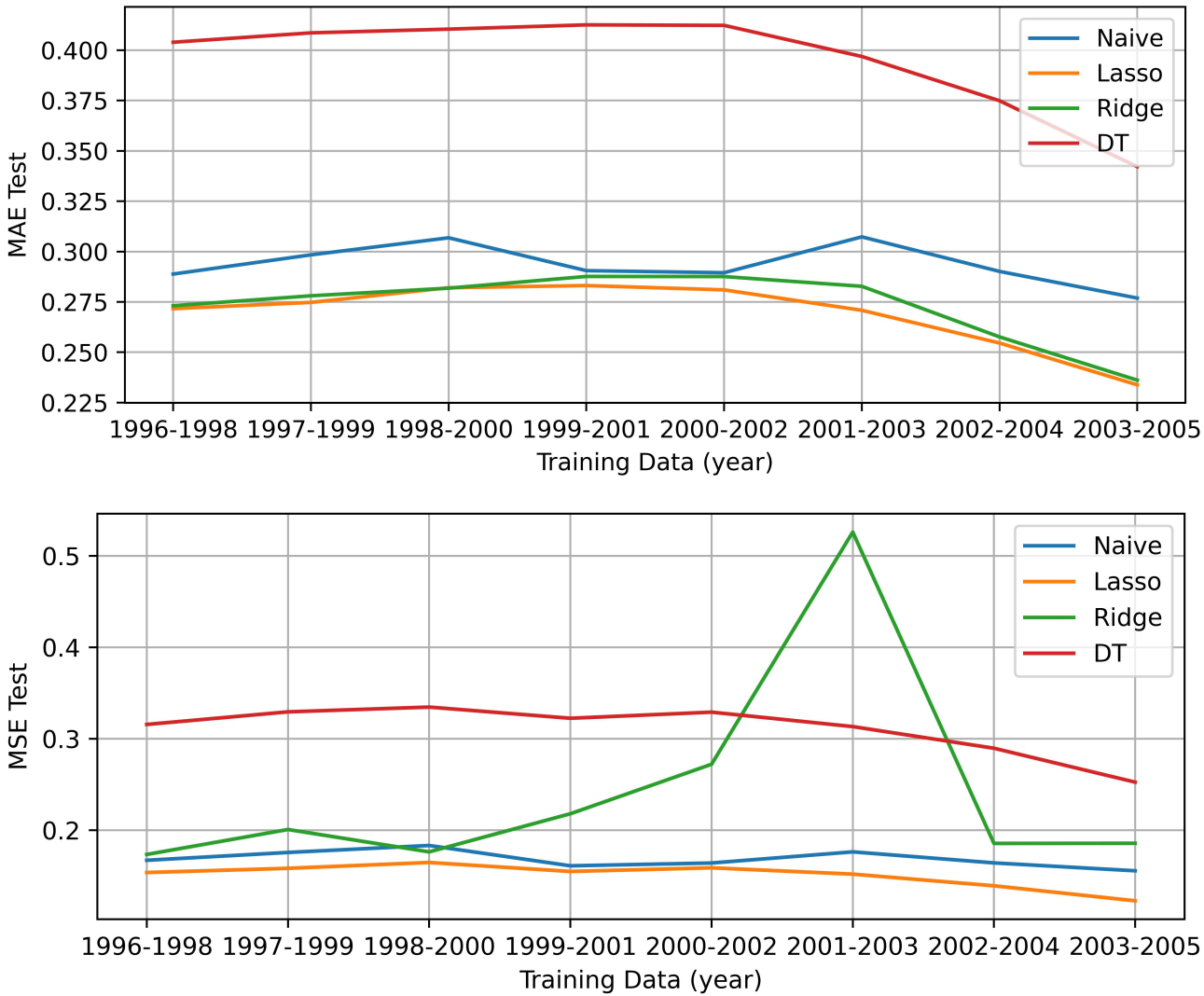

Figure: MAE and MSE for testing 3-year training set.

Result - Error

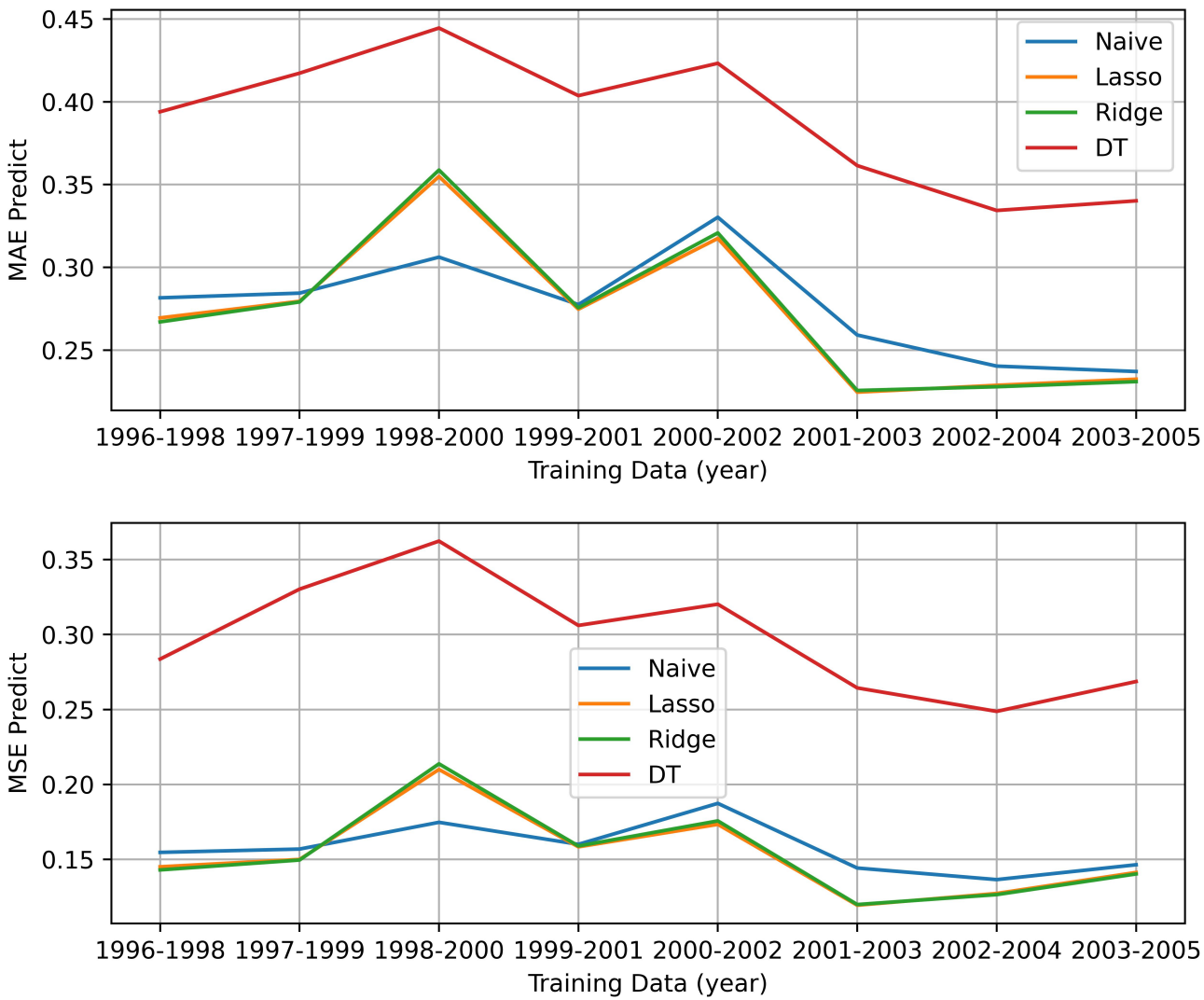

Figure: MAE and MSE for predicting 3-year training set.

Result - Word Features

Year	Increasing Common	Decreasing Common	Different
1996-1997	21.0%	70.0%	0.0%
1997-1998	27.0%	13.0%	0.5%
1998-1999	1.0%	43.0%	1.5%
1999-2000	3.0%	5.0%	0.5%
2000-2001	3.0%	2.0%	0.5%
2001-2002	1.0%	25.0%	1.0%
2002-2003	52.0%	33.0%	0.0%
2003-2004	0.0%	4.0%	2.5%
2004-2005	29.0%	50.0%	0.5%

Table: Word features for lasso regression between different years.

Result - Word Features

Year	Increasing Common	Decreasing Common	Different
1996-1997	11.0%	54.0%	2.5%
1997-1998	29.0%	45.0%	0.5%
1998-1999	0.0%	27.0%	11.5%
1999-2000	0.0%	67.0%	0.5%
2000-2001	2.0%	20.0%	0.5%
2001-2002	29.0%	25.0%	1.0%
2002-2003	24.0%	34.0%	12.5%
2003-2004	2.0%	62.0%	0.5%
2004-2005	34.0%	25.0%	2.5%

Table: Word features for ridge regression between different years.

Result - Word Features

Year	Increasing Common	Decreasing Common	Different
1996-1997	67.0%	9.0%	0.0%
1997-1998	74.0%	11.0%	0.0%
1998-1999	64.0%	24.0%	0.0%
1999-2000	67.0%	10.0%	0.0%
2000-2001	71.0%	32.0%	0.0%
2001-2002	62.0%	8.0%	0.0%
2002-2003	56.0%	8.0%	0.0%
2003-2004	72.0%	15.0%	0.0%
2004-2005	77.0%	35.0%	0.0%

Table: Word features for decision tree between different years.

Result - Word Features

Year	Lasso Ridge	Lasso DT	Ridge DT
1996	68.0%/91.0%/0.0%	8.0%/0.0%/19.0%	6.0%/0.0%/23.5%
1997	62.0%/76.0%/0.5%	27.0%/0.0%/18.5%	39.0%/0.0%/15.5%
1998	91.0%/62.0%/0.0%	33.0%/0.0%/8.0%	36.0%/0.0%/13.5%
1999	19.0%/23.0%/2.5%	6.0%/0.0%/13.5%	7.0%/0.0%/34.0%
2000	35.0%/18.0%/2.5%	18.0%/0.0%/5.5%	11.0%/0.0%/33.5%
2001	22.0%/74.0%/1.5%	7.0%/0.0%/19.0%	34.0%/0.0%/17.0%
2002	95.0%/90.0%/0.0%	24.0%/0.0%/16.0%	26.0%/0.0%/19.0%
2003	59.0%/24.0%/2.5%	22.0%/0.0%/6.5%	16.0%/0.0%/34.0%
2004	56.0%/49.0%/0.5%	8.0%/0.0%/21.0%	13.0%/0.0%/27.0%
2005	76.0%/73.0%/0.0%	24.0%/0.0%/11.5%	23.0%/0.0%/22.0%

Table: Word features between different models (Increasing Common, Decreasing Common, Different).

Conclusion

For forecasting, similar to the previous research[11]:

- More training data can lead to higher accuracy and stablity;
- The Sarbanes–Oxley Act (2002) changes the information/distribution of the words in text;
- The Sarbanes–Oxley Act (2002) makes a higher accuracy in prediction.

[11] Shimon Kogan et al. (2009). "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*, pp. 272–280.

Conclusion

For the words features:

- There exists the similarity between different years;
- The lasso/ridge regression can give the same features;
- The decision tree always choose the same feature, but it seems not good;
- Compare with the features of increasing, the of features decreasing are more similar between years.

Reference I

- Beck, Nico, Jonas Dovern, and Stefanie Vogl (2025). "Mind the naive forecast! a rigorous evaluation of forecasting models for time series with low predictability". In: *Applied Intelligence* 55.6, p. 395.
- Bird et al. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.
- Christensen, Kim, Mathias Siggaard, and Bezirgen Veliyev (2023). "A machine learning approach to volatility forecasting". In: *Journal of Financial Econometrics* 21.5, pp. 1680–1727.
- Kogan, Shimon et al. (2009). "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*, pp. 272–280.
- Lee, Daniel and H Sebastian Seung (2000). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems* 13.

Reference II

- Lee, Daniel D and H Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". In: *nature* 401.6755, pp. 788–791.
- Li, Xiafei, Chao Liang, and Feng Ma (2022). "Forecasting stock market volatility with a large number of predictors: New evidence from the MS-MIDAS-LASSO model". In: *Annals of Operations Research*, pp. 1–40.
- Liang, Chao et al. (2023). "Forecasting China's stock market volatility with shrinkage method: Can Adaptive Lasso select stronger predictors from numerous predictors?" In: *International Journal of Finance & Economics* 28.4, pp. 3689–3699.
- Robertson, Stephen (2004). "Understanding inverse document frequency: on theoretical arguments for IDF". In: *Journal of documentation* 60.5, pp. 503–520.

Reference III


Sparck Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* 28.1, pp. 11–21.