Analysis of Chinese Financial Discourse Based on Topic Clustering and Emotional Evolution

Ruoheng Du

Supervision: Professor Li Guo

2024/04/19

PART 01

Introduction

PART 02

Methodology & Results

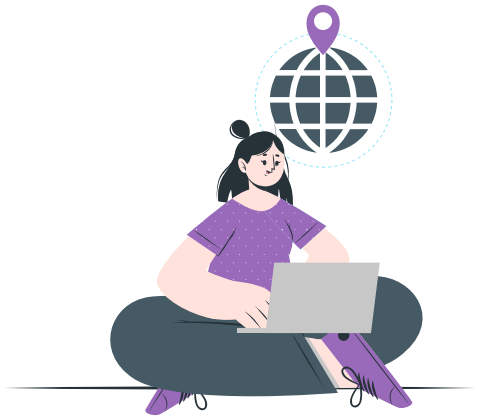
PART 03

Contribution

PART 04

Limitations & Future Directions

Introduction



Social Media

Covid-19

Explore the Influence of Weibo Discussion Topics and Investors' Sentiments from Pre-COVID to Post-COVID

Methodology

Stage I: Topic Modeling

What topics are of interest to investors?

Stage II: Sentiment Analysis

What are the sentiments of investors?

Stage III: Factors for Stock Return Prediction

What factors influence stock returns?

Models derived from Bidirectional Encoder Representations from Transformers (BERT)

BERTopic

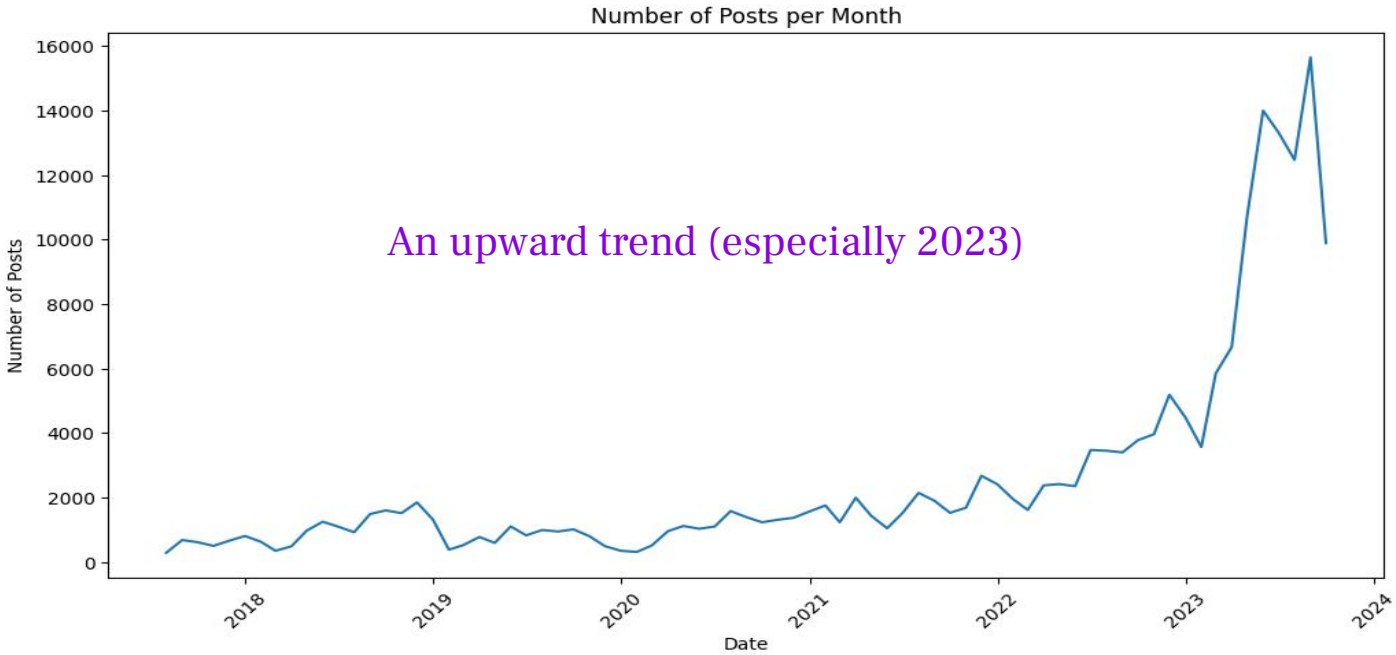
FinBERT

(after Youdao Translation)

Pearson Correlation Coefficient P-value RMSE Reduction

Data Overview

191,460 unique posts from July 13, 2017 to September 24, 2023



An upward trend (especially 2023)

Sample Data

- Remove irrelevant information
  emojis, punctuation, and numbers
- Apply stop list to remove stop words
- Segment sentences into words

date	content	segmented_content	translated_content
2023-01-05 14:49:34	白酒+新能源赛道集体大涨，不是我的菜！今天亏钱已成定局，节后账户三连阳失败！股票	'白酒', '新能源', '赛道', '集体', '大涨', '菜', '亏钱', '已成定局', '节后', '账户', '连阳', '失败'	Liquor new energy track collective rise is not my food today's loss of money is a foregone conclusion after the festival account three consecutive positive failed stocks

RESULT 01

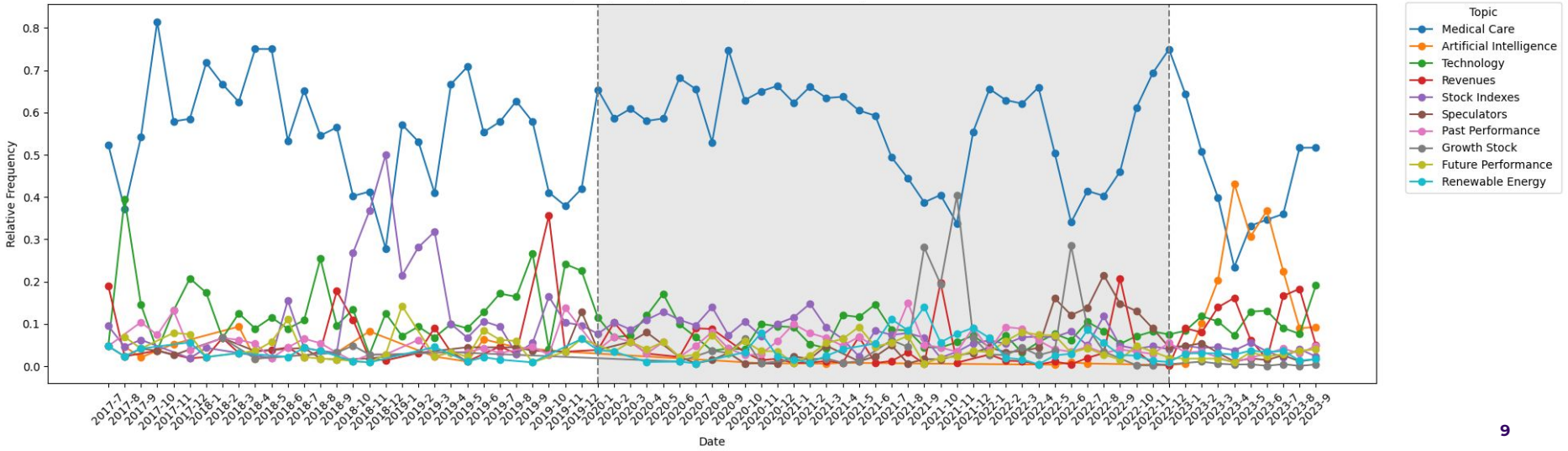
Stage I: Topic Modeling Results

Topic Modeling Results

Topic	Label	Count
0	General Market	117630
1	Medical Care	2071
2	Artificial Intelligence	1568
3	Technology	1439
4	Revenues	1085
5	Stock Indexes	945
6	Speculators	676
7	Past Performance	613
8	Growth Stock	540
9	Future Performance	510
10	Renewable Energy	488

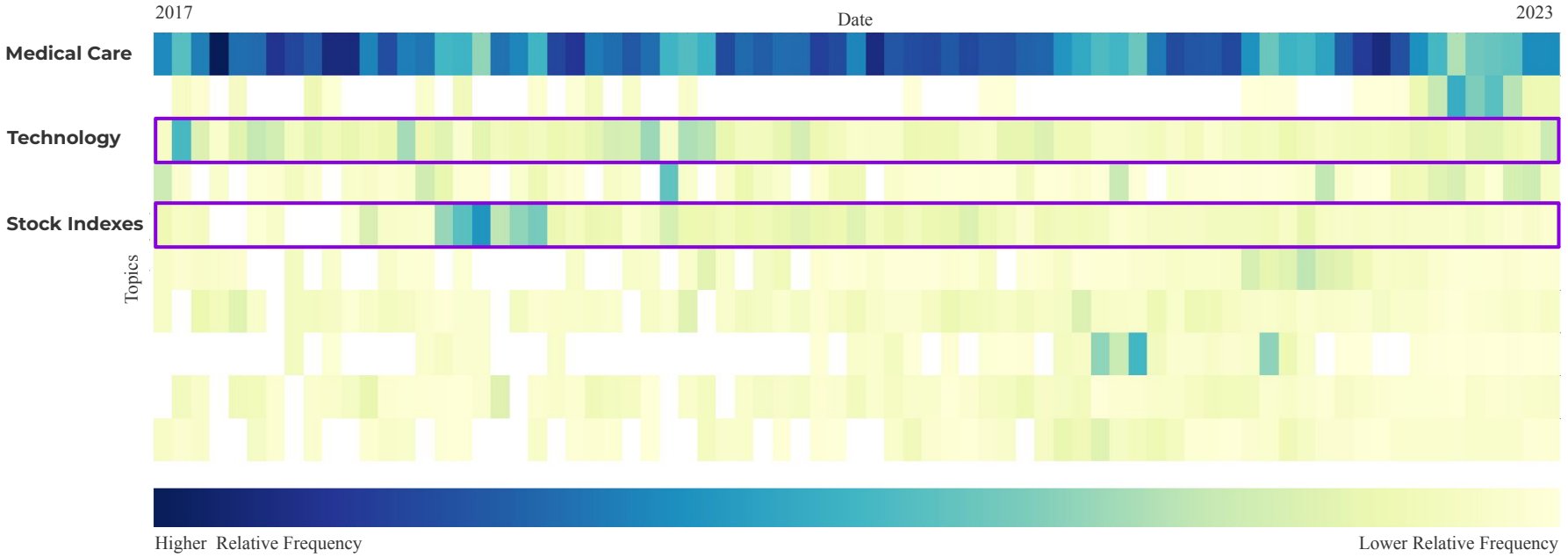
Topic Relative Frequency Over Time I

Medical Care topic almost consistently dominantes other topics (most discussed).



Topic Relative Frequency Over Time II

Technology and Stock Indexes topic repeatedly emerge as the 2nd most discussed.



Topic Relative Frequency Over Time III

Focus changes from Medical Care topic to Artificial Intelligence topic.

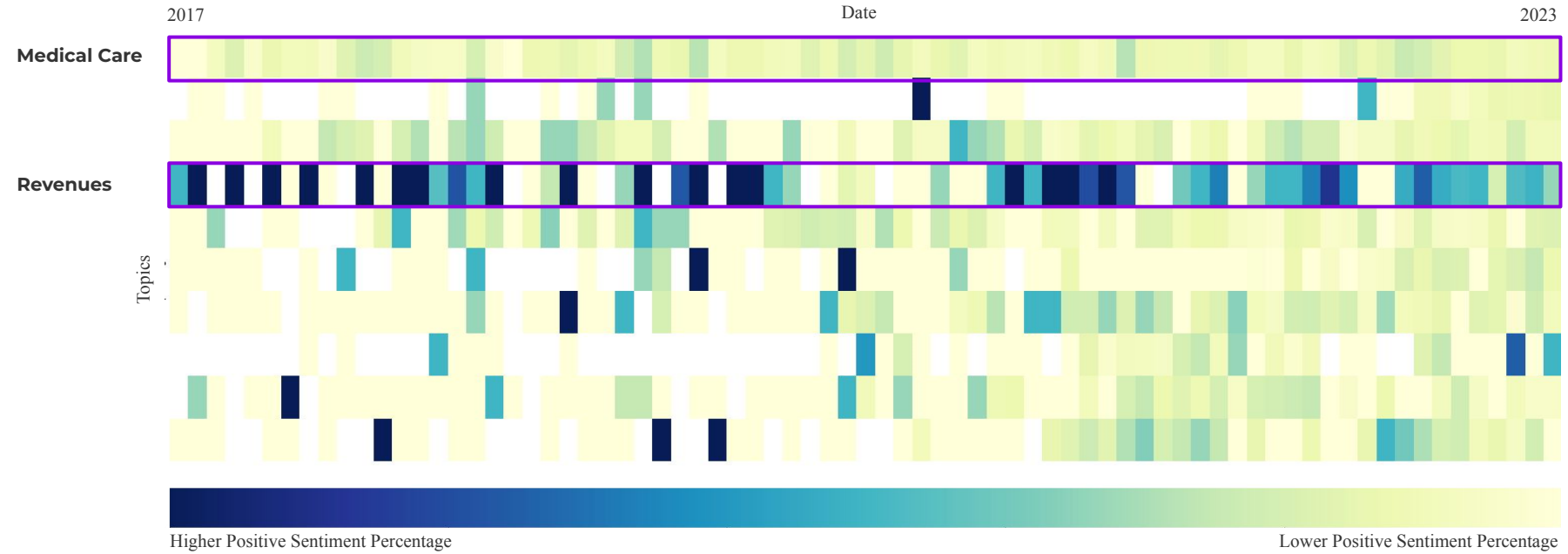
Label	Pre-COVID	During-COVID	Post-COVID	Average
Medical Care	0.56	0.57	0.43	0.52
Artificial Intelligence	0.04	0.01	0.20	0.08
Technology	0.13	0.08	0.11	0.11
Revenues	0.06	0.04	0.11	0.07
Stock Indexes	0.13	0.08	0.04	0.08
Speculators	0.04	0.05	0.03	0.04
Past Performance	0.05	0.06	0.03	0.05
Growth Stock	0.04	0.07	0.00	0.04
Future Performance	0.05	0.04	0.02	0.04
Renewable Energy	0.03	0.04	0.03	0.03

RESULT 02

Stage II: Sentiment Analysis Results

Positive Sentiment Percentage Over Time I

Decreasing trend over time;
Medical Care topic remains low while Revenues topic remains high.



Positive Sentiment Percentage Over Time II

Sentiment shifts: rebounding for Medical Care topic but dropping for Revenues topic.

Label	Pre-COVID	During-COVID	Post-COVID
Medical Care	12.83	12.58	14.25
Artificial Intelligence	17.24	16.67	9.53
Technology	14.35	13.77	11.83
Revenues	64.56	64.77	50.14
Stock Indexes	19.59	10.77	9.36
Speculators	16.67	3.09	14.91
Past Performance	6.67	17.71	14.29
Growth Stock	7.14	8.60	23.08
Future Performance	9.43	11.93	6.47
Renewable Energy	16.67	14.23	17.46

RESULT 03

Stage III: Stock Return Prediction Results

Stock Return Prediction Analysis I

Shanghai Composite Index: Revenues topic stands out.

Labels	Person Correlation	Coefficient P-value	RMSE Reduction
General Market	0.24	0.03	0.12
Medical Care	-0.11	0.35	0.01
Artificial Intelligence	0.20	0.09	0.00
Technology	0.02	0.88	2.86
Revenues	-0.26	0.02	2.54
Stock Indexes	-0.04	0.76	0.02
Speculators	0.10	0.38	0.79
Past Performance	-0.05	0.68	0.40
Growth Stock	0.06	0.60	3.22
Future Performance	0.15	0.20	0.54
Renewable Energy	-0.02	0.87	0.03

Stock Return Prediction Analysis II

Shenzhen Composite Index: Revenues topic stands out.

Labels	Person Correlation	Coefficient P-value	RMSE Reduction
General Market	0.28	0.01	0.06
Medical Care	0.03	0.83	0.27
Artificial Intelligence	0.11	0.34	0.03
Technology	-0.01	0.93	5.33
Revenues	-0.21	0.07	2.60
Stock Indexes	0.00	0.98	0.00
Speculators	0.13	0.27	0.00
Past Performance	-0.04	0.71	0.43
Growth Stock	-0.09	0.45	5.12
Future Performance	0.03	0.77	0.43
Renewable Energy	-0.03	0.77	0.06

Stock Return Prediction Analysis III

CSI 300: Revenues topic stands out.

Labels	Person Correlation	Coefficient P-value	RMSE Reduction
General Market	0.30	0.01	0.60
Medical Care	-0.05	0.66	0.03
Artificial Intelligence	0.19	0.09	0.11
Technology	-0.01	0.93	3.93
Revenues	-0.29	0.01	3.46
Stock Indexes	-0.05	0.65	0.43
Speculators	0.11	0.34	0.21
Past Performance	-0.06	0.59	1.06
Growth Stock	0.03	0.78	4.27
Future Performance	0.10	0.37	0.64
Renewable Energy	-0.05	0.68	0.13

Contribution

1

Enhanced Topic Identification

- Utilize advanced topic modeling technique
- Apply finance-centric sentiment analysis model

2

Insightful Relation Revelation

- Social media sentiment impacts stock returns

3

Impactful Event Understanding

- COVID-19 impacts investor sentiment and stock market dynamics

Limitations & Future Directions

Limitations

- **Dataset Size**

  Limited number of posts per topic

- **Language**

  Machine mistranslation
  - Financial term
  - Colloquial expression

Future Directions

- **Explore Larger Datasets**

  Include more super topics

- **Use domain-specific models**

  Pre-trained models in Chinese financial domain

Thank you!

Special thanks to Professor Li Guo

Professor Christina Wang

Professor Ye Jin

Professor Marti G. Subrahmanyam

References

[1] E. F. Fama, “Efficient capital markets: A review of theory and empirical work,” The Journal of Finance, vol. 25, no. 2, pp. 383–417, 1970.
[2] E. F. Fama, L. Fisher, M. C. Jensen, and R. Roll, “The adjustment of stock prices to new information,” International Economic Review, vol. 10, no. 1, pp. 1–21, Feb. 1969. [Online]. Available: http://ideas.repec.org/a/ier/iecrev/v10y1969i1p1-21.html
[3] E. F. Fama, “Efficient capital markets: Ii,” The Journal of Finance, vol. 46, no. 5, pp. 1575–161, 1991.
[4] ——, “The behavior of stock-market prices,” The Journal of Finance, vol. 38, no. 1, pp. 34–105, 1965. [Online]. Available: http://dx.doi.org/10.2307/2350752
[5] J. Wang, Y. Fan, J. Palacios, Y. Chai, N. Guetta-Jeanrenaud, N. Obradovich, C. Zhou, and S. Zheng, “Global evidence of expressed sentiment alterations during the covid-19 pandemic,” Nature Human Behaviour, vol. 6, no. 3, pp. 349–358, 2022. [Online]. Available: https://doi.org/10.1038/s41562-022-01312-y
[6] R. Xie, S. K. W. Chu, D. K. W. Chiu, and Y. Wang, “Exploring public response to covid-19 on weibo with lda topic modeling and sentiment analysis,” Data and Information Management, vol. 5, no. 1, pp. 86–99, 2021. [Online]. Available: https://doi.org/10.2478/dim-2020-0023
[7] S. Feuerriegel, A. Ratku, and D. Neumann, “Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation,” in 2016 49th Hawaii International Conference on System Sciences (HICSS). Koloa, US: IEEE, New York, 5-8 January 2016, pp. 1072–1081.
[8] E. N. Biktimirov, T. Sokolyk, and A. Ayanso, “Sentiment and hype of business media topics and stock market returns during the covid-19 pandemic,” Journal of Behavioral and Experimental Finance, vol. 31, p. 100542, Sep. 2021. [Online]. Available: https://doi.org/10.1016/j.jbef.2021.100542
[9] M. Baker and J. Wurgler, “Investor sentiment and the cross-section of stock returns,” The Journal of Finance, vol. 61, no. 4, pp. 1645–1680, 2006. [Online]. Available: https://doi.org/10.1111/j.1540-6261.2006.00885.x
[10] G. Zhou, “Measuring investor sentiment,” Annual Review of Financial Economics, vol. 10, pp. 239–259, 2018. [Online]. Available: https://doi.org/10.1146/annurev-financial-110217-022725
[11] Z. McGurk, A. Nowak, and J. C. Hall, “Stock returns and investor sentiment: Textual analysis and social media,” Journal of Economics and Finance, vol. 44, pp. 458–485, 2020. [Online]. Available: https://doi.org/10.1007/s12197-019-09494-4
[12] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” Journal of Computational Science, vol. 2, no. 1, pp. 1–8, Mar. 2011. [Online]. Available: https://doi.org/10.1016/j.jocs.2010.12.007
[13] X. Wang, Z. Xiang, W. Xu, and P. Yuan, “The causal relationship between social media sentiment and stock return: Experimental evidence from an online message forum,” Economics Letters, vol. 216, p. 110598, Jul. 2022. [Online]. Available: https://doi.org/10.1016/j.econlet.2022.110598
[14] M. Grootendorst, “Bertopic: Neural topic modeling with a class-based tf-idf procedure,” 2022.
[15] W. Chen, F. Rabhi, W. Liao, and I. Al-Qudah, “Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: A comparative study,” Electronics, vol. 12, no. 12, p. 2605, 2023. [Online]. Available: https://doi.org/10.3390/electronics12122605
[16] D. Araci, “Finbert: Financial sentiment analysis with pre-trained language models,” 2019.
[17] S. V. Lawrence, COVID-19 and China: A chronology of events (December 2019-January 2020). Congressional Research Service, 2020.
[18] J. Ge, “The covid-19 pandemic in china: from dynamic zero-covid to current policy,” Herz, vol. 48, pp. 226–228, 2023. [Online]. Available: https://doi.org/10.1007/s00059-023-05183-5