Research Proposal

Sector Attention Index in Chinese stock market: Estimating with deep learning methods

Bing HAN

Jan. 4, 2024

Abstract

For Chinese stock market, we define SAI (Sector Attention Index) to quantify the retail investor attention of a specific sector according to the online post volume of the sector’s most representative stocks. We define an abnormal SAI to distinguish extra attention, and a sentiment SAI for sentiment analysis. The text data of post is taken from the most active online stock forum Eastmoney (aka Guba) in mainland China. We rebuild a new sentiment dictionary and design a deep learning model to classify the sentiment tendency in stock posts. We conduct a series of regression analyses to test the predictive power of the SAIs and their correlation with stock returns and trading amount.

Key words: Sector Investor Attention, Sentiment Analysis, Deep Learning, Chinese Stock Market, NLP(natural language processing)

Introduction

The stock market isn’t monolithic. It’s conveniently divided into sectors that group different companies by the types of business they conduct. Major stock indexes, like The S&P 500 (SPX), offer a big-picture view of the entire market, but tracking stock market sectors—such as energy, health care, and technology—can help investors clearly track a given industry’s market performance through time.

In China, the frequency of investor post on stock forums is a good reflection of sector attention. Based on the text analysis results obtained from the stock forums, we can construct a series of indicators to measure the investor attention for different industry sectors, and further study the relationship between these indicators and the market. There is currently no relevant research on measuring sector attention.

The published literature has shown that investor attention have predictive power on stock prices and are correlated with stock returns and trading volume. We will try to examine that these conclusions are also valid in terms of the sector perspective, especially in Chinese stock market.

One the other hand, artificial intelligence technology and alternative data are playing an increasingly important role in financial analysis. For a period of time, researchers focus on how to use big data to study investor attention and sentiment, which is contained in social platforms like Twitter or machine readable news on financial websites. However, researches based on stock forum data are relatively limited, especially for the Chinese stock market. Our research will explore this field and use more advanced NLP & AI-based methods to process unstructured text data.

Let's focus on a latest news. A few days ago, NeuroIPS (Conference and Workshop on Neural Information Processing Systems) 2023 released its award-winning paper, with the Time Test Award awarded to the NeuroIPS paper "Distributed Representations of Words and Phrases and Their Composition" from ten years ago. This work introduces the groundbreaking word embedding technology word2vec, demonstrating the ability to learn from a large amount of unstructured text and driving the arrival of a new era of natural language processing (NLP).

In reality, advanced NLP technologies with great potential such as word2vec have not yet been widely adopted in the financial field. In my research, I will introduce how to use deep neural network models to optimize text processing results. We have used some advanced technologies, such as phrase detection based on likelihood ratio testing and neural networks based on word embedding. The purpose of doing so is to capture a richer dictionary and to obtain more accurate sentiment classification results.

Literature Review

Antweiler and Frank (2004) studied posts on Yahoo Finance online forums and found that attention metrics measured by the number of posts can effectively predict stock returns and market volatility. Post sentiment divergence is positively correlated with stock trading volume during the same period.

Zhi et al. (2011) obtained the weekly Google search index for individual stocks and directly measured investor attention using the search frequency of stocks. Research has shown that using search indices can more timely measure investor attention, and an increase in search indices can predict stock price increases in the next two weeks and stock price reversals within a year.

As Zhang et al. (2016) found, emotions on Twitter have a certain predictive effect on the Dow Jones Industrial Average. When emotions on Twitter are expressed strongly, such as showing a large number of emotional factors such as hope and concern, the Dow Jones Industrial Average will decrease the next day.

Sprenger et al. (2014) studied forums dedicated to discussing the stock market on Twitter, extracting keywords to conduct in-depth research on individual stocks and major company issues. The results showed a correlation between the sentiment of Twitter articles and stock returns and trading volume.

Li et al. (2018) examined the extent to which stock microblog messages (i.e., tweets) are related to financial market indicators and the mechanism leading to efficient aggregation of information. They collected more than 1.2 million messages related to S&P 100 companies and analyzed the data on both a daily and a 15-min basis. They found that the sentiment of messages is positively affected with daily abnormal stock returns and the message volume could predict 15-min follow-up returns, trading volume, and volatility.

Chen C P et al. (2018) first used CNN to calculate the sentiment of Chinese retail investors and compared the predictive performance of CNN and SVM models. Their research found that the prediction accuracy of CNN is roughly equivalent to SVM, but the CNN model is more decisive in classification.

Chen Z et al. (2023) utilized the Robinhood investor data and the Google Search Volume Index to measure active retail investor attention, and found that active retail investor attention is impacted by recent stock returns and more significant for larger stocks.

Chen S et al. (2020) proposed a new proxy to measure asymmetric attention of retailers with signed (positive, negative and neutral attitude) posts published on online stock message board of Chinese A-share market. It shows that the proxy for asymmetric attention is significant and positive related to volatility asymmetry. Furthermore, we find that negative information arrivals can induce higher volatility asymmetry, and asymmetric attention which acts as a mediator incorporates more negative information flows into market, which then triggers high asymmetric volatility. Moreover, this proxy is an independent variable from idiosyncratic financial leverage, but its influence on asymmetric volatility increases with market systematical risks.

Dixon's (2022) book covers fundamentals in using alternative-data for investing and sets out best practices for pre-processing and modeling with alternative-data.

Klaas' (2019) book explores advances in machine learning and how to put them to work in financial industries. It gives clear explanation and expert discussion of how machine learning works, with an emphasis on financial applications and a coverage of advanced machine learning approaches including neural networks, GANs, and reinforcement learning.

Mikolov et al. (2013) introduced a continuous Skip-gram model, an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships.

Lai et al. (2022) examined whether Google search volume index (GSVI), a proxy of investor attention, can predict the excess returns and abnormal trading volumes of TPEx 50 index constituents. It also explores the motive underlying GSVI based on positive or negative shocks to stock prices.

Booker et al. (2018) used posts on the investor-focused StockTwits social media network to generate new insights regarding investor disagreement, disclosure processing costs, and trading volume around earnings announcements. They found that both pre-announcement disagreement and increases in disagreement around an earnings announcement are positively associated with trading volume.

Chen J et al. (2022) proposed an investor attention index based on proxies in the literature and find that it predicts the stock market risk premium significantly, whereas every proxy individually has little predictive power. The predictive power stems primarily from the reversal of temporary price pressure and from the stronger forecasting ability for high-variance stocks.

Loughran and McDonald (2011) develop an alternative negative word list, along with five other word lists, that better reflect tone in financial text. They link the word lists to 10-K filing returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings.

Fama and MacBeth (1973) tested the relationship between average return and risk for New York Stock Exchange common stocks. The theoretical basis of the tests is the "two-parameter" portfolio model and models of market equilibrium derived from the two-parameter portfolio model.

Questions

How do we measure investor attention (especially the retail investor) of stock market sectors in China?

- For Chinese stock market, we define SAI (Sector Attention Index) to quantify the investor attention of a specific sector according to the online post volume of the sector’s most representative stocks. We concern how to get an abnormal SAI to distinguish which sectors are suddenly receiving extra attention in the market. We also concern about the sentiment feature in the SAI.

What textual data is available in active stock market forums to support our research?

- The retail investor posts in online stock forums are a good reflection of attention to specific sectors. We collect text data from Eastmoney (aka Guba), the largest and most active stock communication platform in Chinese Mainland.

What advanced AI methods can we use in the analysis of text data?

- We use Part-of-speech tagging method to rebuild a new sentiment dictionary and design a deep learning model to classify the sentiment tendency in stock posts.

What is the correlation between the trading data (e.g. stock returns and trading amount) and the investor attention indicators we defined?

- We will conduct a series of regression analyses to test the predictive power of sector attention index and its correlation with stock returns and trading amount.

Methods

Key Variable Definitions

1. Sector Attention Index (SAI)

We will define an indicator to measure investors' attention of different sector indices and we call it the Sector Attention Index (SAI). For sector *i*

$$SAI_i = \sum_{k=1}^{N} V_{i,k}$$

Where *N* is the number of selected stocks which can represent the sector to the greatest extent, and $V_{i,k}$ is the post volume of stock *k* within a single time cycle (e.g. last 30 days).

2. Abnormal SAI

The investor attention between different sectors cannot be compared horizontally, because there is a certain positive correlation between the number of posts and the size of the sector. Some sectors with fewer constituent stocks and lower total market value will definitely not have as many posts as larger sectors, so we cannot use absolute values for comparison. An feasible approach is to use abnormal SAI to distinguish which sectors are suddenly receiving extra attention in the market.

We define abnormal SAI as

$$abnormalSAI_t = \log(SAI_t) - \log[Med(SAI_{t-1}, \ldots, SAI_{t-r})]$$

where log (SAI_t) is the logarithm of SAI during week *t*, and log [$Med(SAI_{t-1}, \ldots, SAI_{t-r})$] is the logarithm of the median value of SAI during the prior *r* weeks.

Zhi et al. (2011) have confirmed in their paper that this way of definition is efficient. Intuitively, the median over a longer time window captures the "normal" level of attention in a way that is robust to recent jumps. It also has the advantage that time trends and other low-frequency seasonalities are removed. A large positive abnormal SAI clearly represents a surge in investor attention and can be compared across stocks in the cross-section.

3. sentiment SAI

We will create a classification task to categorize investor posts into two categories. Each post is marked as a positive sentiment (marked as 1) or a negative sentiment (marked as 0). Based on the classification results, we can calculate the sentiment SAI as follows

$$sentimentSAI_i = \frac{1}{N}\sum_{k=1}^{N}\frac{V_{i,k}^{positive}}{V_{i,k}}$$

Where $V_{i,k}{}^{positive}$ is the positive sentiment post volume of stock *k,* sector *i.*

Data

1. Text data

The text data of post is taken from the online stock forum Eastmoney (aka Guba). This is the largest and most active stock communication platform in Chinese Mainland, where retail investors post and comment. These posts and comments are often very short, averaging less than 20 words.

The posts to collect include ***the most representative stocks*** in main industry sectors. The text collected based on these stocks is used as a corpus for the sector. This is done because investor discussion is not active under the sector index theme, while very active under the representative stocks'. Meanwhile, compared to index forums, posts on individual stock forums express strong buying and selling preferences. Therefore, it is very suitable for the investor attention analysis. We will collect data for a certain period of time (usually not less than one year).

2. Market trading Data

The market data we collect includes major indicators such as daily opening and closing prices, as well as trading volume and amount. The collection targets include the sector indices (SI), and representative stocks of each SI. We will use SI category and daily data (Table 1.) from RoyalFlush Network Technology, a Chinese financial data service provider.

Table 1. Sector index data demo, published by RoyalFlush Network Technology

Sector	Date	Open	High	Low	Close	Volume	Amount
881101	20201013	3299.52	3305.41	3261.98	3283.49	382250000	3850680000
881102	20201013	2796.46	2850.65	2775.58	2846.61	301790000	6608240000
881103	20201013	3965.01	3996.09	3935.66	3989.45	651560000	10429000000
881104	20201013	10602.2	10906.8	10588.9	10835.1	140700000	2852140000

The most representative stocks

Although the number of posts on a single stock is not always considerable, a group of constituent stocks can effectively compensate for this disadvantage and is qualified to represent the entire sector. As of now, the RoyalFlush has released over 80 industry & sector indices. Most sectors have more than 30 constituent stocks and they are often stocks with high weights and a concentration of investors.

We will not collect data from all sectors, but rather select a portion from them. The number of selected stocks N for each sector will also be balanced based on the data time period. More importantly, we will measure the constituent stocks that are most relevant to the sector index history (e.g. having the highest correlation coefficient value), but cannot choose them based on their weights, in order to avoid the impact of abnormal fluctuations in individual stocks.

Post data from website

We will develop a web crawler program to collect retail investor posts in Eastmoney's stock forum and insert them into database. For sector and component information, we also need to execute a separate program to collect sector list and constituent stock list. After we combine the two sets of information, we get the raw dataset as shown in Table 2. The first column Code represents a ticker symbol.

Table 2. Raw dataset demo, the Eastmoney investor post

Code	Sector	Time	Content	Author
600999	Security	2020/1/13 13:25:**	招行都撑不住，大盘真不行 China Merchants Bank cannot hold on, the overall market is really not good.	股友 2Ev***
002230	Computer application	2020/2/13 13:42:**	三天内会创新高的 It will reach a new high within three days	股友 pnb***

Text data preprocessing

We need delete posts that do not meet the requirements, including those that are suspected of advertising and those that are too short. We adopt Jieba tokenizer for word segmentation. Jieba tokenizer is a highly effective Chinese segmentation device that supports multiple segmentation modes.

We will try a text phrase detection tool (***Appendix 1***) based on likelihood-ratio test (a statistical method), which can detect common phrases in retail investor comments without manually setting grammar rules, and enhance the accuracy of word segmentation. The purpose of doing this is to expand our segmentation result to make the text analysis more robust.

In order to adapt to the embedding layer of AI model, We make all sequences equal length by cutting them or adding zeros at the end. We will fix the length of all posts to 40 characters. This step can be achieved through the methods provided by Keras, an open-source artificial neural network library written in Python.

Classification task for sentiment SAI

1. A new, targeted sentiment dictionary

Most scholars choose the existing general sentiment dictionaries and word banks as reference to conduct their researches, which leads to the lack of pertinence to the specific context (e.g., stock market comments are high in technical terms, social media is high in slang and emojis). A sentiment dictionary should be built specifically for stock market posts or comments. We can remove general sentiment words and use Part-of-speech (POS) tagging method to preserve names, acts and adjectives as much as possible. By sorting the frequency of word appearances, we can pick up alternative new sentiment words and score them with Amazon Mechanical Turk (Mturk). Of course, this work can also be done manually.

2. Deep learning model

This research builds a BiLSTM-CNN-Attention sentiment analysis model to classify the sentiment tendency in stock posts. The structure of this model may probably include word embedding layer, BiLSTM layer, CNN layer, attention layer and output layer. The final version will depend on the situation during the research process. Refer to ***Appendix 2*** for more details about the model. Appropriate pre-trained resources for Chinese NLP are also necessary.

Regression method

Fama and MacBeth (1973) conducted a two-step regression method, a practical way for measuring how correctly these risk factors explain asset or portfolio returns. The first step was to perform least squares regression on the explanatory and dependent variables of each stage to obtain estimated parameters, treating each parameter as a sample value of the overall parameter; The second step is to take the average of all parameters in the first step and calculate the estimated parameters of the overall data. We will use this method for our empirical testing.

We will refer to the empirical methods of Loughran and McDonald (2011) to select the explained variables, including excess returns, abnormal trading amount, volatility of returns and etc.. Descriptive statistics need to be performed on the variables of the regression model, including sample size, mean, standard deviation, minimum and maximum values. ***The regression results of the SAIs on trading amount, volatility, and unexpected returns should show significant correlation***.

Use out of sample data for robustness testing. Through this test, we can demonstrate the effectiveness of our method in preprocessing and sentiment classification of non sample text data. The significant correlation presented should also be consistent with the sample data.

Further Discussion

1. Regret Index

We found that the posts and comments related to individual stocks reflect more of the psychological biases of investors described in behavioral finance. For example, there are a large number of samples with words related to “regret”, whether it is regret not buying or regret not selling. We also want to test if there is a significant correlation between regret sentiment and stock returns, and to what extent does the increase in investor attention translate into actual trading (buy or sell).

Table 3. Examples of investor posts related to “regret”

Code	Sector	Time	Content	Author	Regret for
600999	Security	2020/3/25 11:31:**	16.88 没卖后悔啊 I regret not selling at 16.88 yuan	股友 JVz***	Not selling
600999	Security	2020/7/6 10:25:**	周五 3 个点跑的，后悔 Sold at 3 points on Friday, regretting	股友 1r2***	Sold out early

Reference

[1] Antweiler, W. , & Frank, M. Z. . (2004). Is all that talk just noise? the information content of internet stock message boards. Journal of Finance, 59(3), 1259-1294.

[2] Zhi, D. A. , Engelberg, J. , & Gao, P. . (2011). In search of attention. Journal of Finance, 66(5), 1461-1499.

[3] Zhang, Y. , Song, W. , Shen, D. , & Zhang, W. . (2016). Market reaction to internet news: information diffusion and price pressure. Economic Modelling, 56, 43-49.

[4] Sprenger, T. O. , Tumasjan, A. , Sandner, P. G. , & Welpe, I. M. . (2014). Tweets and trades: the information content of stock microblogs. European Financial Management, 20(5), 926–957.

[5] Li, T. , Van Dalen, J. , & Van Rees, P. J. . (2018). More than just noise? examining the information content of stock microblogs on financial markets. Journal of Information Technology, 33(1), 50-69.

[6] Chen, C. P., Tseng, T. H., & Yang, T. H. (2018, June). Sentiment Analysis on Social Network: Using Emoticon Characteristics for Twitter Polarity Classification. In International Journal of Computational Linguistics & Chinese Language Processing, Volume 23, Number 1, June 2018.

[7] Chen, Z., & Craig, K. A. (2023). Active attention, retail investor base, and stock returns. Journal of Behavioral and Experimental Finance, 100820.

[8] Chen, S., Zhang, W., Feng, X., & Xiong, X. (2020). Asymmetry of retail investors' attention and asymmetric volatility: Evidence from China. Finance Research Letters, 36, 101334.

[9] Matthew Dixon (2022) The Book of Alternative Data: A Guide for Investors, Traders and Risk Managers, Quantitative Finance, 22:8, 1427-1428, DOI: 10.1080/14697688.2022.2078736

[10] Klaas, J. (2019). Machine learning for finance: principles and practice for financial insiders. Packt Publishing Ltd.

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

[12] Lai, H. H., Chang, T. P., Hu, C. H., & Chou, P. C. (2022). Can google search volume index predict the returns and trading volumes of stocks in a retail investor dominant market. Cogent Economics & Finance, 10(1), 2014640.

[13] Booker, A. , Curtis, A. , & Richardson, V. J. . (2018). Investor disagreement, disclosure processing costs, and trading volume: evidence from investors who interact on social media. Social Science Electronic Publishing.

[14] Chen, J., Tang, G., Yao, J., & Zhou, G. (2022). Investor attention and stock returns. Journal of Financial and Quantitative Analysis, 57(2), 455-484.

[15] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10 - Ks. The Journal of finance, 66(1), 35-65.

[16] Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of political economy, 81(3), 607-636.

Appendix 1

A text phrase detection method and device based on likelihood ratio test
[Unapplied patent]

Introduction

This invention belongs to the field of text phrase detection in NPL. It adopts text segmentation technology and likelihood ratio test, avoids manual intervention in the detection process as far as possible, and improves the applicability and reliability of the phrase detection.

The prior art solution (by searching for the most relevant patents) attempts to presuppose a part-of-speech dependency rule that applies to most phrases by traversing all alternative word combinations to find a match. If the corpus is limited to a specific field, and there are many specialized words, then this should be the best approach. However, this default rule requires a lot of human intervention, and needs to be updated frequently as the corpus develops, which is difficult to fully achieve in practical applications, and it is not suitable for tweets, blogs, and news articles. In addition, when calculating the word formation probability of word combination, it is often necessary to manually preset the model threshold, and its size will also affect the quality of the final phrase detection, so the reliability is not high.

Methodology

The function of word segmentation is to extract all words contained in the text based on rules, and different word segmentation devices have different segmentation rules. Jieba tokenizer is currently a highly effective Chinese segmentation device that supports multiple segmentation modes. This present adopts the precise mode of the jieba tokenizer, which is implemented based on Python.

The likelihood ratio is defined as the ratio of the maximum value of the likelihood function under constrained conditions to the maximum value of the likelihood function under unconstrained conditions. The procedure of the likelihood ratio test is: for a given pair of words, the method tests two hypotheses on the observational data set. The first hypothesis (null hypothesis) states that the occurrence of word 1 is independent of word 2. The second hypothesis (the alternative hypothesis) states that seeing word 1 changes the likelihood of seeing word 2. If we accept the alternative hypothesis, it means that these two words can form a common phrase. Which hypothesis holds is determined by calculating the frequency of words actually observed in the corpus.

The general processing flow is to segment the input text into words by jieba tokenizer, filters the stop word, assembles the n-gram, and calculates the likelihood ratio. The n-grams are sorted according to the calculated result of likelihood ratio, and the n-grams with the smallest likelihood ratio are selected as the final detection result.

Appendix 2

BiLSTM-CNN-Attention model

This research builds a BiLSTM-CNN-Attention sentiment analysis model to extract the sentiment tendency in stock posts. The structure of this model includes

1. Word embedding layer: The word vector of the text is constructed by pre-trained word2vec model, and the sentences after word segmentation are mapped into low-dimensional and dense vectors, with each word corresponding to a vector. The generated word vectors contain semantic information, which is conducive to further feature extraction in the next layer.

2. BiLSTM layer: The Bidirectional Long Short Term Memory (BiLSTM) model uses the stacking of two layers of LSTM to break free from the limitation that the model can only predict the output of the next time based on the temporal information of the previous time, and can better combine context for output.

3. CNN layer: The CNN (recurrent neural network) performs convolution operations on the output of BiLSTM at each time step, further extracting text features, and adding max pooling after convolution to prevent overfitting, reducing parameters and computational complexity.

4. Attention layer: The weight matrix of the attention mechanism is obtained by assigning different probability weights and multiplying them with the CNN output vector, and then counting them through the softmax function to obtain the value of the weight matrix.

5. Output layer: The output layer maps the results through the sigmoid function to obtain the results of sentiment classification.

Available pre-trained resource

Pre-trained models or algorithms publicly disclosed by some AI giants (e.g. Tencent and Baidu) will help us improve the efficiency of NLP tasks for Chinese text. We will choose some of these including kinds of 1) Pre-trained models which can complete typical sentiment analysis tasks such as Sentence-level Sentiment Classification and Aspect-level Sentiment Classification. 2) Embedding corpora for Chinese words and phrases which provide finite-dimensional vector representations, a.k.a. embeddings, for Chinese words. We will introduce these pre-trained models to build our word embedding layer and BiLSTM layer, after finetuning them to suit our stock analysis scenarios.