Comparison of the Prediction Effect of the Garch-x Model on Sse 50 Volatility Based on the Baidu and Douyin Search Indices

Hongjin Chen1,[[1]](#footnote-1), Ruiyu Liu2, and Zhuoxin Wu3

1School of Shipping Economics and Management, Dalian Maritime University, Dalian, China

2School of International Economics and Trade, Central University of Finance and Economics, Beijing, China

3SILC Business School, Shanghai University, Shanghai, China

Abstract. With the rapid development of financial markets and the advancement of information technology, the application value of alternative data in economic research has become increasingly prominent. This paper takes the Baidu search index and Douyin search index as research objects to explore their predictive effect on the volatility of the Shanghai Composite 50 Index. Based on data from 2021 to 2024, the study constructs a GARCH (1,1) basic model and a GARCH-X model that taking the search index as an exogenous variable, the performance of the two search indices in volatility prediction is compared and analyzed. The results show that after adding the search index, the prediction accuracy of the model has improved, among which the effect of the Baidu search index is more significant. Further analysis shows that the Baidu search index data has small fluctuations and strong synchronisation with market sentiment, and can reflect market information more stably; while the Douyin search index data fluctuates violently and has lags, and the prediction effect is relatively weak. This study fills the research gap of Douyin search index in the financial field and provides a reference for future multi-source data fusion and model optimization.

1. Introduction

In China's financial market, with the continuous advancement of capital market reform, the full implementation of the registration system, and the deepening of opening up to the outside world, the market complexity has increased significantly [1]. As the "barometer" of the Chinese stock market, the Shanghai Composite 50 Index not only reflects the overall performance of large-cap blue-chip stocks but also has a profound impact on the flow of market funds and investor confidence. In this context, in-depth exploration of the intrinsic connection between the search index and the Shanghai Composite 50 Index will help investors to grasp the changes in market sentiment promptly and optimize investment decisions [2]. It can also provide regulatory authorities with a new perspective for market public opinion monitoring and enhance risk warning and market regulation capabilities. With the in-depth development of the financial market and the iterative upgrading of information technology, data sources have shown an explosive growth trend. In addition to traditional financial data, alternative data is gradually becoming a new focus of research and practice in the financial field with its unique information value. At present, the competition in the global financial market is becoming increasingly fierce, and investors are more and more eager to accurately predict market trends. The search index in alternative data has opened up a new path for financial market research because it can capture the behavior and emotions of market participants in real time [3,4].

At present, search index data has shown great potential in financial market research and applications, especially in stock price turning point prediction and volatility prediction. According to the research of Shen et al., Baidu Search Index can predict the changes in stock prices the next day, and investor attention is negatively correlated with stock prices. This study supplements the existing literature on the predictability of stock returns in China [5,6]. Fang et al. used the GARCH model and found that after adding Baidu search index data, the prediction accuracy of China's stock market volatility was improved, proving that search index data contains relevant information about stock volatility [7]. In the study of time series, Chen et al. found that the combination of ARIMA and a multivariate linear regression model can more accurately predict the number of traffic accident deaths than using only the ARIMA model. This research idea can be transferred to the stock field [8]. However, current research on search indices mostly focuses on Google Trends and the Baidu search index. As a short video platform with more than 1 billion monthly active users and extremely fast information dissemination speed, Douyin's corresponding Douyin search index is still a blank area in financial market research [9,10].

This paper obtains the relevant data of the Shanghai Stock Exchange 50 Index from 2021 to 2024 from the Massive Arithmetic, Baidu Index, and CSMAR platforms. In the study, the GARCH model is first used to compare the prediction effects of the two indexes on the volatility of the Shanghai Stock Exchange 50 Index, and the reasons for the different prediction effects are analyzed based on the index construction method, etc.; secondly, a combined model of GRACH and multivariate linear regression is constructed to explore whether it can improve the prediction accuracy of volatility. The standard deviation and variance of the Baidu search index are smaller, and the volatility of the data is lower. The maximum and minimum values of the Douyin search index differ too much, and the data volatility is very large.

The research results of this paper can not only fill the gap in the Douyin index in the financial market research, but also provide important references for subsequent researchers when facing multivariate data selection by comparing the prediction performance of different search indexes, and help promote the innovative development of financial market research and practice.

1. Methodology
   1. Data source

The daily return data of the Shanghai Composite 50 Index from January 4, 2021, to December 31, 2024 (969 trading days) comes from the CSMAR platform. The Baidu search index (keyword "Shanghai Composite 50") data from December 31, 2020, to December 31, 2024, comes from the Baidu Index platform. From December 31, 2020, to December 31, 2024, the Douyin search index (keyword "Shanghai Composite 50") data comes from the Ocean Engine platform. In this article, return is used to represent the daily return of the Shanghai Composite 50 Index, dy is used to describe the Douyin search index value with the keyword "Shanghai Composite 50", and bd is used to represent the Baidu search index value with the keyword "Shanghai Composite 50".

Calculation method of daily return rate of Shanghai Composite 50 Index:

(1)

Where:

represents the return rate of index n on day t

represents the closing index of index n on day t

represents the closing index of index n on day t-1 (i.e., the previous day)

* 1. Descriptive statistics and correlation analysis

Based on the data obtained, this paper conducts descriptive statistics on Douyin search index, Baidu search index, and the daily return rate of the Shanghai Composite 50 Index. The results are shown in Table 1. And a normality test is performed on them; the results are shown in Table 2.

Table 1. Descriptive statistics

|  |  |  |
| --- | --- | --- |
|  | Statistical indicators | Value |
| return | Mean | -0.24711 |
| Median | -0.059700 |
| Variance | 1.347 |
| Minimum | -6.0433 |
| Maximum | 7.1689 |
| dy | Mean | 524.77 |
| Median | 266.00 |
| Variance | 1952589.524 |
| Minimum | 44 |
| Maximum | 27654 |
| bd | Mean | 1931.17 |
| Median | 1796.00 |
| Variance | 283060.462 |
| Minimum | 1221 |
| Maximum | 7166 |

Table 2. Normality test

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smith test | | | Shapiro-Wilk test | | |
|  | Statistics | Degrees of freedom | Significance | Statistics | Degrees of freedom | Significance |
| return | 0.063 | 969 | 0.000 | 0.956 | 969 | 0.000 |
| dy | 0.367 | 969 | 0.000 | 0.215 | 969 | 0.000 |
| bd | 0.134 | 969 | 0.000 | 0.782 | 969 | 0.000 |

From Table 1, it can be seen that the Baidu search index data has a small range, a small variance, while the Douyin search data has a significant range and a large variance. After observing the Douyin search index, its early data sparseness may reflect the stage characteristics of user search habits. Therefore, the data quality of the Douyin search index is poorer than that of the Baidu search index. According to the significance test results in Table 2, all three datasets obey the normal distribution.

* 1. Correlation analysis

This paper uses the Spearman correlation coefficient to conduct a correlation analysis on the Baidu search index, the Douyin search index, and the daily return of the Shanghai Composite 50 Index. By comparing the correlation coefficients, analyzed how much information related to the return of the SSE 50 Index is contained in the two search indices. The results are shown in Table 3.

Table 3. Correlation

|  |  |  |  |
| --- | --- | --- | --- |
|  | return | dy | bd |
| return | 1.000 | 0.018 | 0.119 |
| dy | 0.018 | 1.000 | 0.578 |
| bd | 0.119 | 0.578 | 1.000 |

From Table 3, it can be seen that the correlation coefficient between return and bd is 0.119 and significant, and there is a certain positive correlation between the two; the correlation coefficient between return and dy is only 0.018, which is not significant, and the correlation is weak. This shows that under the same period, the Douyin search index contains less information about the return rate of the Shanghai Composite 50 Index, while the Baidu search index contains certain information. The Douyin search index may contain more information after the nth lag. Therefore, in Section 4, this paper determines the lag order of the two indices added to the model.

1. Model construction and prediction
   1. Model setting

The idea of this paper is to use the Garch model for basic prediction first, and then add two alternative data, Baidu search index and Douyin search index, as exogenous variables to construct the Garch-X model, then perform fitting and prediction, as shown in Table 4. Basic GARCH model form:

= (2)

Substituting the Shanghai Composite 50 Index yield data into Formula 2, the corresponding can reflect the volatility of the Shanghai Composite 50 Index return, and will be predicted later.

Then, the Garch-X model is used to predict the Baidu search index and the Douyin search index, which are exogenous variables:

= (3)

Table 4. GARCH model-related variables

|  |  |
| --- | --- |
| Variable | Meaning |
|  | Conditional variance of the Shanghai Composite 50 Index return at time t |
|  | Constant term |
|  | Past residual square term, is its coefficient |
|  | Past conditional variance, is its coefficient |
|  | The past exogenous variable term, is its coefficient |
|  | Maximum lag order of the past residual square term |
|  | The maximum lag order of the conditional variance term in the past |
|  | The maximum lag order of the exogenous variable term in the past |

* 1. Selection of lag order

According to the AIC rule, p=1, q=1 are determined. That is, the Garch (1, 1) model, as shown in Formula 4:

= (4)

The Baidu search index and the Douyin search index with lags of 1-4, respectively, are added to the GARCH model, and the value and significance of γ are tested. The results are shown in Table 5.

Table 5. γ value and its significance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | No lag | First-order lag | Second-order lag | Third-order lag | Fourth-order lag |
| bd | 0.000329 | 0.000172 | 0.000172 | 0.000112 | 0.00008 |
| dy | -0.000008 | 0.000077 | 0.00004 | 0.00004 | 0.000022 |

From Table 4, it can be seen that the Baidu search index is not lagged, and the Douyin search index is lagged by one order are best choice to add to the model.

* 1. Add to index

Garch-X model of Baidu search index added with the same period of daily return of the Shanghai Composite 50 Index:

= (5)

Garch-X model of Douyin index added with lagged first order:

= (6)

* 1. Rolling forecast of volatility

This paper uses the daily return of Shanghai Composite 50 Index from January 4, 2021 to December 29, 2023 (a total of 730 trading days) as the training set, and the daily return of Shanghai Composite 50 Index from January 2, 2024 to December 31, 2024 (a total of 242 trading days) as the validation set, and perform rolling forecasts.

Base Model represents the basic Garch model that only substitutes the Shanghai Stock Exchange 50 yield data, BD Model represents the Garch-X model with the addition of Baidu search index as an exogenous variable, and DY Model represents the Garch-X model that adds Douyin search index as an exogenous variable.

1. Result evaluation and analysis

This study uses the mean square error (MSE) as an indicator to evaluate the fitting effect of each model. The smaller the MSE value, the smaller the deviation between the model prediction value and the actual value, that is, the higher the prediction accuracy of the model, as shown in Table 6.

Table 6. Comparison of prediction effects

|  |  |  |  |
| --- | --- | --- | --- |
|  | Base Model | DY Model | BD Model |
| MSE | 1.303312 | 1.274724 | 1.230449 |

In terms of data characteristics, Baidu search index data shows the characteristics of stable mean and small standard deviation, relatively gentle data fluctuation, strong synchronization with market sentiment, and can reflect market-related information more stably. In contrast, the range of Douyin search index is more significant, the standard deviation is larger, the data fluctuates violently, and the early data has a more obvious sparse problem, which makes the stability of the information transmitted in the model prediction poor, thereby weakening the accuracy of the prediction.

From the perspective of user behavior, the user groups on the Baidu platform mostly search for financial information. The data generated by their search behavior can more directly reflect investors' immediate attention to the SSE 50 Index and has a closer relationship with market fluctuations. However, the user behavior of the Douyin platform is characterized by entertainment, and the search behavior is not mainly centered on financial investment, which leads to a relatively weak internal connection between its search data and the market fluctuations of the SSE 50 Index.

At the model adaptability level, through the study of model construction and variable relationships, it is found that the Baidu search index has a significant impact on the volatility of the Shanghai Stock Exchange 50 Index in the current period and can provide effective information for the model promptly. The Douyin search index needs to lag for one period to play a better role in the model. This phenomenon shows that the market information it reflects has a certain delay in the process of transmitting and influencing market fluctuations, which reduces the model's efficiency in capturing and utilizing information in real time.

In December 2024, the prediction effect of the DY Model dropped sharply (as shown in Figure 1). The following reasons may exist: Structural changes in the market: At the end of the year, the market was affected by policy adjustments, large-scale capital flows, and other factors, and the market volatility pattern changed suddenly. The market sentiment and related information reflected by the Douyin search index failed to adapt to this structural change promptly, and could not effectively provide accurate information for the model, increasing the model prediction deviation.

Data collection deviation: The Douyin platform may have updated its algorithm during this period, or an abnormality occurred in the data collection interface, which made the search index data unable to accurately reflect the user's real search behavior and the actual correlation between the market, causing the correlation between the Douyin search index and the market volatility of the Shanghai Composite 50 Index to break, and then the prediction effect of the DY Model suddenly deteriorated.

1. Conclusion

Based on the relevant data of the Shanghai Composite 50 Index from 2021 to 2024, this study uses the GARCH and GARCH-X models to compare and analyze the prediction effects of Baidu and Douyin search indexes on the volatility of the Shanghai Composite 50 Index. The study confirmed that the introduction of a search index can improve the model's prediction ability. Baidu search index performed better, with an MSE of 1.230449, an increase of 7.29% over the basic model; Douyin search index MSE was 1.274724, an increase of 2.86%. From the three dimensions of data characteristics, user behavior, and model adaptability, it was found that Baidu search index data was stable, closely related to market fluctuations, and had a significant impact in the current period, while Douyin search index had problems such as large data fluctuations and delayed information transmission. Its prediction effect dropped sharply in December 2024, which may be due to changes in market structure and data collection deviations. This study fills the gap in Douyin search index research in the financial market and provides a reference for subsequent exploration. Future research can be deepened and expanded from both data and models. At the data level, expand the sample range, include more financial indexes and long-term data, and enhance the universality of the conclusions; introduce alternative data such as social media comments and industry news to enrich model input information. At the model level, deep learning technology is used to explore the potential relationship between the Douyin search index and market fluctuations, and optimize data preprocessing; a hybrid model is constructed to break through the limitations of traditional models and improve the ability to capture complex data; the performance of different models in extreme market environments is compared, and their stability is evaluated to provide financial market participants with more accurate prediction tools and decision-making basis.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. D. Shen, Y. Zhang, X. Xiong, W. Zhang, Baidu index and predictability of Chinese stock returns. Financ. Innov. **3**, 4 (2017)
2. J. Fang, G. Gozgor, C.-K.M. Lau, Z. Lu, The impact of Baidu Index sentiment on the volatility of China's stock markets. Financ. Res. Lett. **32**, 101099 (2020)
3. J. Gu, H. Yang, J. Zhang, X. Zhang, Analysis of pricing deviations of stock indices and individual stocks based on keyword search popularity and trading strategy suggestions. Tsinghua PBC Sch. Econ. MBA (2023)
4. Y. Wang, S. Wang, K.K. Lai, A new fuzzy support vector machine for evaluating credit risk. Inf. Sci. **180**, 1523-1536 (2010)
5. R. Dash, P.K. Dash, Stock market prediction using the Firefly Algorithm with Deep Learning. Procedia Comput. Sci. **114**, 55-63 (2017)
6. D. Kumar, S. Murugan, Performance analysis of stock price prediction using an artificial neural network. Procedia Comput. Sci. **54**, 462-470 (2015)
7. Y. Huang, Prediction of the Stock Price of Shanghai Securities Composite Index by Using Time-series Mode. Proc. 2022 Int. Conf. Econ. Bus. Manage. (2022)
8. Z. Jiang, D. Xu, J. Liang, Deep reinforcement learning for stock trading. arXiv:**1808**, 09940 (2018)
9. A. Altan, S. Karasu, S. Bekiros, Digital currency forecasting with chaotic meta-heuristic bio-inspired signal processing techniques. Expert Syst. Appl. **182**, 115215 (2021)
10. H. Ni, S. Wang, P. Cheng, A hybrid approach for stock trend prediction based on tweets embedding and historical prices. World Wide Web **24**, 849-868 (2021)

1. Corresponding author: hongjin\_chen@dlmu.edu.cn [↑](#footnote-ref-1)