A Novel Approach to Economic Sentiment Index Based on Structured and Unstructured Data

Yuqing Duan

yuqing.duan@auckland.ac.nz

University of Auckland, 12 Grafto RD, Auckland Central, Auckland, 1010, New Zealand

Abstract. This study proposes a new approach to analysing economic sentiment indexes and stock market movements: combining structured financial indicators with qualitative sentiment analysis and topic theme modelling in news texts. The study firstly points out the limitations of traditional economic analyses relying only on structured data. For example, the cyclicality of China's A-share market over the past decade - peaking in 2015 and then declining sharply - reflects market volatility that is difficult to fully capture with traditional methods. In this paper, five models are compared and it can be obtained that DNN is able to identify deep nonlinear patterns in data. Therefore it performs well in analysing both structured and unstructured data. In conclusion, this study highlights the shortcomings of traditional economic analyses that focus only on structured data and points out the importance of integrating unstructured data (e.g., market sentiment and news analyses) for a more accurate understanding of financial health and market dynamics.

Keywords: Economic Sentiment Index, News Analysis, Financial Indicators, Deep Neural Network

1 Introduction

Economic cycle fluctuations are cyclical changes in economic growth and recession. This performance is characterised by leading indicators that predict economic expansion or contraction [1]. During periods of economic boom, markets are usually marked by prosperity and growth; conversely, during periods of recession, there may be a tendency towards contraction and recession. Behind these economic fluctuations, the Economic Sentiment Index (ESI) becomes an essential tool for assessing the overall robustness of the economy. Traditionally, most analyses have used time series analysis. However, this approach has shown some limitations in complex economic phenomena. Although multiple regression analysis has become a new trend in structured data analysis, more is needed to fully reflect the multidimensional complexity of the economy by relying on structured data alone. As shown in **Figure 1. [2]** below, the economic cycles that have affected China's A-share market over the past decade can be observed, as represented by the Shanghai Stock Exchange Composite Index. The apparent peak reached in 2015, followed by a precipitous fall, encapsulates the inherent volatility of the stock market - ostensibly reflecting the inflation and contraction of the economy. It not only demonstrates the marked volatility of markets but also highlights the need to anticipate unforeseen market changes.

ICBBEM 2024, March 29-31, Wuhan, People's Republic of China

DOI 10.4108/eai.29-3-2024.2347401

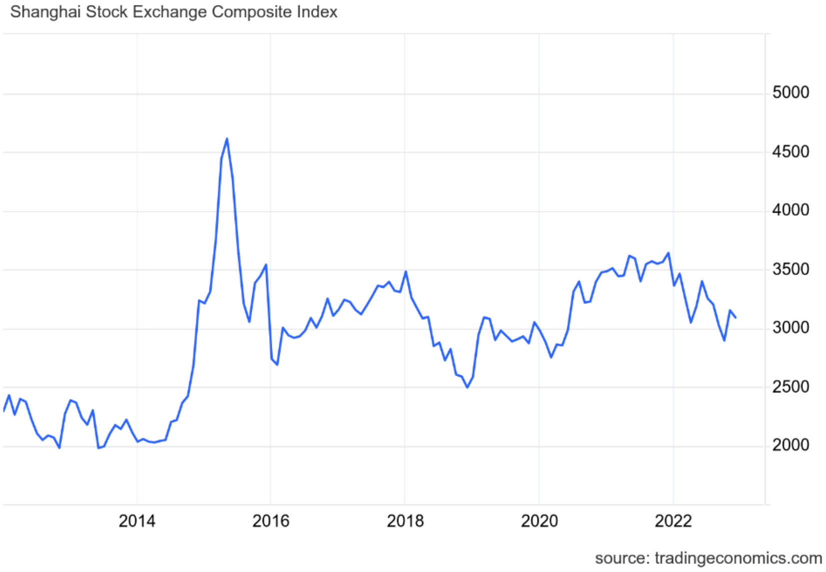

Figure 1. China Shanghai Composite Stock Market Index from 2012 to 2023.

In order to overcome this challenge, this study adopts an approach that combines structured and unstructured data. This paper considers more than 20 structured data, such as financial indicators, etc., and news texts with unstructured data are integrated. The focus is to do sentiment analysis on the text data by combining Bert's pre-training model and then refining the critical information on market sentiment using topic modeling (e.g., the LDA model) for classification.

2 Literature Review

The Economic Sentiment Index (ESI) has existed since the early 20th century. The core of ESI is the integration of different economic statistics [3] [4]. Traditionally, ESI is constructed from explicit data such as corporate financial and macroeconomic data. However, this traditional approach may not fully grasp the market's complexities. As an illustration of the limitations inherent in economic indicators, Figure 2. [5] delineates China's GDP growth from 2012 to 2023. The graph showed a decrease from 7.85% in 2012 to just 2.8% in 2020, indicating severe economic challenges. This drop in GDP growth could only be partially explained by structured data alone. Also, when looking at the stock market (as depicted in Figure 1. [2]), standard economic measures, including GDP growth, fail to shed light on the 2015 stock market boom. This unusual stock market activity suggests that traditional indicators miss some key factors. As a result, there's increasing interest in a more comprehensive approach that considers unstructured data, such as market sentiment and news, to get a fuller picture.

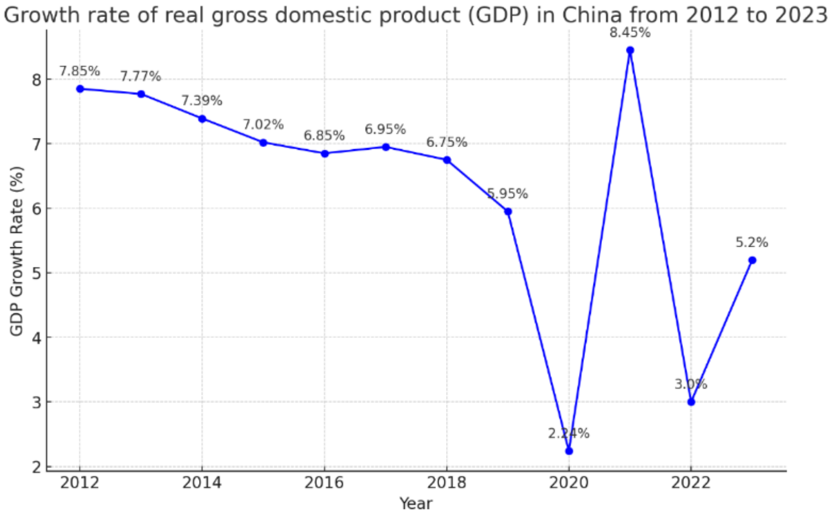

Figure 2. Growth Rate of Real Gross Domestic Product (GDP) in China from 2012 to 2023.

Researchers increasingly focus on diverse data types, including the sentiments and themes in news reports, to better understand market dynamics. These insights are crucial in grasping investor behaviour and market trends [6] [7]. Several techniques can be useful when analysing unstructured data such as news articles. For example, the Latent Dirichlet Allocation (LDA) model is effective for exploring large amounts of digital news to spot trends and patterns. The sentiments and themes found in news pieces are helpful in creating economic sentiment indices. Additionally, how financial news uses language has been linked to stock market movements. Studies have shown that positive wording often goes together with market rises, while negative wording might predict market falls. [8].

In the stock market, sentiment analysis of news aims to discover the opinions and emotions hidden in the text. In this regard, some studies confirm that discussions and comments on social media can provide immediate feedback about market trends, company events, or macroeconomic policies [8]. Meanwhile, a study investigated the impact of news articles' sentiment on the stock market. They found that good or bad news affects investor sentiment and behaviour, changing market performance [7]. The remaining studies have compared news articles about companies with their stock performance. In conclusion, most of these studies focused on the sentiment of news headlines to determine stock price changes, emphasising the importance of sentiment analysis in understanding stock market ups and downs.

Classical tools such as SARIMA and linear regression are commonly used to analyse time series data. For example, SARIMA is used to study the factors affecting the sales of consumer goods [9]. On the other hand, linear regression, based on the principle of least squares, looks for relationships between different factors [10]. These models are known for their simplicity and ease of understanding, which can help economists make predictions based on past data. However, they only sometimes capture changing markets.

Academic research has been conducted on using machine learning in economic forecasting. Modern models can handle large and complex datasets, and machine learning can quickly acquire reliable data trends. Scholars, therefore, believe that machine learning has an advantage over traditional methods in dealing with complex data [11]. Research in this field ranges from simple regression to complex neural networks. These studies emphasised that machine learning models can improve economic forecasts' accuracy and effectiveness [12]. Recently, studies have applied extreme learning machines (ELMs) to forecast gross domestic product (GDP). ELMs can learn efficiently and process complex data quickly, which makes them well-suited for economic forecasting [9]. Moreover, recently discussed neural network models can give more accurate predictions. They do this by weighing different factors differently, demonstrating the flexibility and adaptability of machine learning in predicting economic trends [13]. Another development is a Bert-based approach for finding sentiment and critical entities in online financial texts and social media. Tests have shown that, in some cases, this method outperforms well-known methods such as SVM, LR, NBM, and even BERT. It is very good at analysing financial sentiment and picking out essential entities from the dataset.

However, combining advanced machine learning models (e.g., LightGBM, DNN, and LSTM) with traditional economic models takes work. While these models are helpful for big data, they are built differently than conventional economic models. This leads to challenges in making them work together, such as figuring out how to blend them properly [14]. Furthermore, choosing a suitable model for a particular data can be tricky. Much depends on what the data looks like, its complexity, and what you try to predict.

3 Methodology

3.1 Dataset

Data are obtained from the CSMAR database and the Ministry of Finance of the People's Republic of China [15]. The average method is used to integrate the variables for all companies in the a-share market annually, effectively reducing the complexity of the data and minimising the impact of short-term market fluctuations.

The structured dataset covers 11 years (2012-2022) and includes data from four major stock indices: the Shanghai Stock Exchange (SSE) a-share, the Shenzhen Stock Exchange (SZSE) a-share, the Growth Enterprise Market (GEM), and the SSE Star Market, representing a total of 4,772 companies. The critical financial variables in Table 1 include company economic and performance indicators. This is complemented by data on China's treasury rates from 2011 to 2023, coded as 1 (rising) and 0 (falling) to represent monetary policy fluctuations. The unstructured dataset contains a collection of news items for the same period. The selection criterion for the news data is to select the largest file each day, totalling 4,278 news items, including information on publishing platforms, titles, and contents.

3.2 Analysis Process

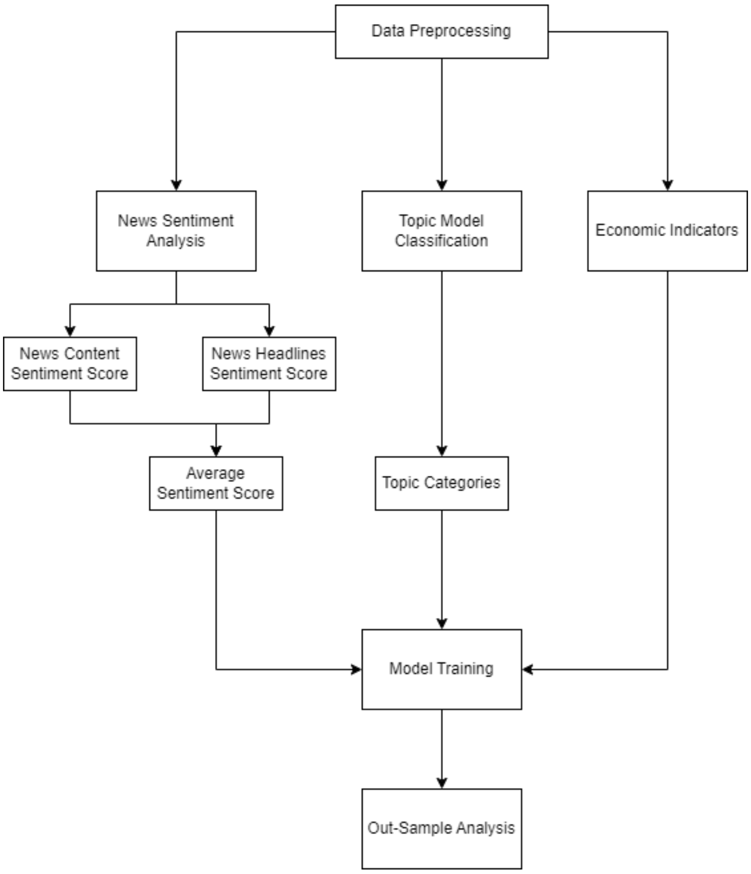

Figure 3. Analysis Process.

The analysis flow is shown in **Figure 3**:

Data Preprocessing includes null value imputation, removal of stop words, and elimination of special characters. For categorical data, label encoding is applied, and numerical data are normalized using a Minmax normalization technique:

$$\text{Normalized Value} = \frac{\text{Value} - \text{Min}(\text{Value})}{\text{Max}(\text{Value}) - \text{Min}(\text{Value})} \tag{1}$$

News Sentiment Analysis: Sentiment analysis of news headlines and content was performed using a Bert pre-trained model. The model is capable of extracting positive sentiment scores from the text. Finally, the sentiment scores will be averaged across all news headlines and content:

$$\text{Sentiment Score} = \frac{\sum_{i=1}^{n} \text{Sentiment}_i}{n} \tag{2}$$

Topic Model Classification: News content is classified using the LDA topic model, preset to five categories. The Latent Dirichlet Allocation (LDA) model assigns probabilities of topics to each news and words to each topic. They are Policy and Regulation (0), Market Dynamics (1), Industry Innovations (2), Company Earnings and Results (3), International Impact (4)

The model's output can be represented as:

$$\mathrm{P(Topic|News)} = \frac{\mathrm{P(News|Topic)} * \mathrm{P(Topic)}}{\mathrm{P(News)}} \tag{3}$$

Economic Sentiment Index (ESI) Construction

The dependent variable, the Economic Sentiment Index, represents the A-share market closing index. The independent variables include the financial indicators specified earlier. The first set of analyses focuses only on A-share market data. In contrast, the second set of analyses combines this data with the results of news and thematic studies.

Models: Various models, including SARIMA, Linear Regression, LightGBM, DNN, and LSTM, were deployed to capture and analyze the intricate features and interrelations in the data.

For Regression Models:

Firstly, just A-share market variables and bonds.

$$\begin{aligned}&\text{Economic Sentiment Index}_t \\ &= \beta_0 + \beta_1 * \mathrm{Lev}_t + \beta_2 * \mathrm{Size}_t + \beta_3 \mathrm{ROA}_t + \ \dots + \ \beta_n * \mathrm{rdspendsumratio}_t + \varepsilon_t\end{aligned} \tag{4}$$

β_0 is the intercept

$\beta_1, \beta_2, \dots, \beta_n \gamma$ are coefficients to be estimated.

ε_t: is the error term.

Second Model: Combines A-share market variables and bonds with news and thematic analysis data.

$$\begin{aligned}&\text{Economic Sentiment Index}_t \\ &= \beta_0 + \beta_1 * \mathrm{Lev}_t + \beta_2 * \mathrm{Size}_t + \beta_3 \mathrm{ROA}_t + \ \dots + \ \beta_n * \mathrm{rdspendsumratio}_t + \varepsilon_t \\ &\quad + \beta_{n+1} * \text{Sentiment Score}_t + \beta_{n+2} * \text{Topic Probability}_t\end{aligned} \tag{5}$$

Sentiment Score_t : represents the average sentiment score from the news.

$\text{Topic Probability}_t$: is the probability of news topics at time t.

Both models aim to analyze the impact of each independent variable on the ESI through the estimation and testing of coefficients (β and γ) using standard statistical methods, ensuring model validity and reliability.

Metrics: The evaluation of each model's effectiveness, considering the concentrated nature of the constructed ESI (values less than 1), is based on mean square error (MSE), mean absolute error (MAE), and mean fundamental percentage error (MAPE).

4 Results and Discussion

The dataset was divided into two parts: one using only structured data (21 features) and the other mixing structured data with unstructured data (news data) (23 features).

4.1 Analysis of News Texts

Positive sentiment scores in news texts are usually associated with good economic performance, whereas negative sentiment may signal a recession. Results of the news analysis can be learned from **Table 2**: During 2012-2014, the Chinese economy achieved moderate growth, and the government shifted the economy from export-led to consumption-led. Therefore, the economic sentiment was cautiously optimistic [16]. However, in 2015-2016, China's economy was challenged by slowing growth and stock market turmoil, leading to a more negative sentiment [17]. The 2017-2018 period saw signs of stabilization, with improvements in technology and consumer goods, and the advancement of the Belt and Road Initiative may have boosted sentiment [18]. 2019 would see new challenges posed by the U.S.-China trade tensions, with the U.S. government shifting from an export-led to a consumer-led economy. In 2020-2021, sentiment is more volatile due to the impact of the Covid-19 Epidemic on the economy but turns positive as the economy gradually recovers. By 2022, economic sentiment will remain cautiously optimistic despite challenges such as supply chain disruptions, reflecting global uncertainty and the resilience of the Chinese economy.

Table 2. Results of News Analysis.

Year	Content Sentiment Score	Headline Sentiment Score	Average Score	LDA Category
2012	0.52	0.5	0.51	0
2013	0.54	0.48	0.51	1
2014	0.52	0.56	0.54	2
2015	0.44	0.51	0.48	1
2016	0.49	0.47	0.48	0
2017	0.58	0.52	0.55	3
2018	0.5	0.52	0.51	4
2019	0.48	0.46	0.47	4
2020	0.32	0.36	0.34	2
2021	0.52	0.56	0.54	1
2022	0.58	0.56	0.57	0

Using the LDA topic model to categorise news content provides insights into market dynamics across economic activities. This can be seen in the fourth column of **Table 2**:

In 2012, the focus was on policies and regulations, reflecting the government's strategies to regulate and supervise the market; in 2013 and 2015, market dynamics became the core topic, covering volatility, growth, or downward trends in the stock market. In 2014 and 2020, industrial innovations became the primary news topic as technology and industries developed rapidly. Policy and regulation are again in the news in 2016 and 2022, faced with the post-epidemic period of market instability and global economic recovery. In 2017, corporate earnings and performance are in the spotlight, reflecting economic activity at the corporate level. In 2018 and 2019, international influences became a major news category, reflecting the significant impact of global events on China's stock market. In 2021, market dynamics were again in the spotlight as the market gradually recovered from the epidemic. These annual themes reflect significant trends and macroeconomic changes in China's stock market.

4.2 Comparison of Model Performance

For the model performance, it can be seen in Table 3. The mean square error (MSE) and mean absolute error (MAE) of all the models are relatively close, but the LSTM model has the worst performance, with its MSE and MAE much higher than the other models. This suggests that the LSTM model may not be as effective as the other models when dealing with situations containing only structured data. In contrast, DNN performs best in terms of MSE but is higher in terms of Mean Absolute Percentage Error (MAPE), which suggests that DNN is too aggressive in some of its predictions, resulting in more significant percentage errors.

When structured and unstructured data are combined, as can be seen in Table 4, the performance of all models improves especially the DNN model, which significantly outperforms the other models on all three metrics, indicating that DNN is more effective in dealing with complex data structures. In addition, LightGBM showed significant improvement in MAPE after combining the data, which may reveal its better adaptation to unstructured data.

The performance of the SARIMA model remained the same in both cases, suggesting that it is insensitive to the increase in unstructured data, which may be because SARIMA itself is more suited to dealing with time-series data rather than complex situations that contain multiple types of data. When combining data, the linear regression model improves on all metrics, suggesting that it can handle richer data sets effectively. Still, it remains higher on MAPE than SARIMA and LightGBM, implying that its predictions are less accurate than those models.

Table 3. Adoption of Structured Data.

Model	MSE	MAE	MAPE (%)
SARIMA	0.17	0.37	27.15
LinearRegression	0.17	0.34	38.78
LightGBM	0.17	0.34	35.78
DNN	0.12	0.37	41.85
LSTM	0.31	0.51	62.73

Table 4. Combining Structured and Unstructured (News).

Model	MSE	MAE	MAPE (%)
SARIMA	0.17	0.37	27.15
LinearRegression	0.14	0.31	36.6
LightGBM	0.17	0.34	16.99
DNN	0.05	0.21	28.16
LSTM	0.22	0.37	41.85

The advantage of deep neural network (DNN) models in analysing structured data stems mainly from their ability to powerful feature learning and pattern recognition. Structured data, such as tables, database records, or explicitly formatted documents, often contain a wealth of quantitative information such as numbers, text, etc. DNN models are able to learn the complex structures and underlying relationships in these data through the cascading of multiple layers of neurons. Each layer of neurons transforms and combines the input data, gradually abstracting it so that deep, non-linear features can be captured.

5 Conclusion

This study reveals the limitations of traditional economic analysis of structured data and highlights the potential of integrating unstructured data to provide an understanding of market dynamics. This research introduces a novel method for grasping economic sentiment and stock market behaviour. It does this by merging financial indicators with qualitative insights from analysing news articles' sentiments and topics. What's interesting is that machine learning models, especially deep neural networks (DNNs) that can handle complex data, seem to better understand the many layers of market sentiment and economic performance.

To sum up, the study points out the value of using a diverse analysis approach. It combines time-tested economic models with cutting-edge machine learning methods. This blended approach could make economic predictions more accurate and give investors and policymakers new insights, helping them navigate the intricate world of financial markets.

Appendix

Table 1. Variable Names and Description.

Variable Name	Description
Year	The key variable for time series analysis, is used to observe trends over time.
size	Typically measured by the company's market capitalization or total assets.
nnindnme	An unspecified variable that needs to be defined for better analysis.
lev	Leverage ratio, reflecting the degree of a company's financial leverage.

roa	Return on Assets, a measure of a company's asset profitability.
roe	Return on Equity indicates the profitability of a company to its shareholders.
cashflow	Cash Flow Ratio, assessing the health of a company's cash flow.
growth	Revenue Growth Rate, reflecting the trend in a company's income growth.
agrowth	Total Asset Growth Rate measures the growth in a company's assets.
ato	Asset Turnover Ratio, reflecting the operational efficiency of a company's assets.
bm	The book-to-market ratio is used to assess the proportion of a company's market value to its book value.
rec	Proportion of Receivables The proportion of a company's receivables to its total assets.
inv	Inventory Proportion indicates the proportion of inventory in a company's total assets.
fixed	Fixed Asset Ratio is the proportion of fixed assets in total assets.
intangible	The intangible asset ratio is the proportion of intangible assets to total assets.
mfee	Management Expense Ratio is the proportion of management expenses.
pe	Price-to-Earnings Ratio, reflecting the ratio of stock price to earnings per share.
pb	The price-to-book ratio is the ratio of stock price to net assets per share.
investor	Proportion of Institutional Investors, the stake of institutional investors in a company.
rdspendsum	Total R&D Expenditure: total amount a company invests in research and development.
rdspendsuratio	R&D Expenditure Ratio, the proportion of R&D spending to business income.
Government Bond Rate	As a macroeconomic indicator, it reflects the government's debt level and fiscal health, significantly impacting the stock market.

References

[1] Silgoner, M. A.: The Economic Sentiment Indicator. Vol. 2007, No. 2, pp. 199–215. Journal of Business Cycle Measurement and Analysis, (2008)

[2] Trading Economics. “China Shanghai Composite Stock Market Index | 2023 | Data | Chart | Calendar,” Tradingeconomics.com, (2023). https://tradingeconomics.com/china/stock-market

[3] Khramov, V. and Ridings Lee, J.: The Economic Performance Index (EPI): an Intuitive Indicator for Assessing a Country’s Economic Performance Dynamics in an Historical Perspective. Vol. 13, No. 214, p. 1. IMF Working Papers, (2013)

[4] Mamais, K. and Karvelas, K.: Feeling good, as a guide to performance: the impact of economic sentiment in financial market performance for Germany. Vol. 52, No. 41, pp. 4529–4541. Applied Economics, (2020)

[5] Textor, C.: China GDP Growth Rate 2011-2023 | Statistic," Statista, Oct. 20, 2023. https://www.statista.com/statistics/263616/gross-domestic-product-gdp-growth-rate-in-china/

[6] Nemes, L and Kiss. A.: Prediction of stock values changes using sentiment analysis of stock news headlines. pp. 1–20. Journal of Information and Telecommunication, (2021)

[7] Kim, J. Seo, J. Lee, M and Seok J.: Stock Price Prediction Through the Sentimental Analysis of News Articles. 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), (2019)

[8] TETLOCK, P. C. SAAR-TSECHANSKY, M. and MACSKASSY, S.: More Than Words: Quantifying Language to Measure Firms' Fundamentals. Vol. 63, No. 3, pp. 1437–1467. The Journal of Finance, (2008)

[9] Sokolov-Mladenović, S. Milovančević, M. Mladenović, I. and Alizamir, M.: Economic growth forecasting by artificial neural network with extreme learning machine based on trade, import and export parameters. Vol. 65, pp. 43–45. Computers in Human Behavior, (2016)

[10] Geisser, S.: Introduction to Fisher (1922) On the Mathematical Foundations of Theoretical Statistics. pp. 1–10. Springer series in statistics, (1992)

[11] Paruchuri, H.: Conceptualization of Machine Learning in Economic Forecasting. Vol. 11, No. 2, pp. 51–58. Asian Business Review, (2021)

[12] Shobana, G. and Umamaheswari, K.: Forecasting by Machine Learning Techniques and Econometrics: A Review. IEEE Xplore, (2021)

[13] Zhi Yu Zheng: Economic forecasting based on neural network with weight learning and local connection. International Conference on Neural Networks, Information, and Communication Engineering (NNICE 2022), (2022)

[14] Ribeiro, M. T. Singh, S. and Guestrin, C.: Why Should I Trust You?. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, (2016)

[15] "ChinaBond-- ChinaBond Yield Curves," yield.chinabond.com.cn. https://yield.chinabond.com.cn/

[16] Center for Macroeconomic Research of Xiamen University, "China's Macroeconomic Outlook Quarterly Forecast and Analysis Report," 2016. Accessed: Feb. 21, 2024. [Online]. Available: https://content.e-bookshelf.de/media/reading/L-8490295-5107ecffc7.pdf

[17] Cashin, P. Mohaddes, K. and Raissi, M.: China's Slowdown and Global Financial Market Volatility: Is World Growth Losing Out?. Vol. 16, No. 63, p. 1. IMF Working Papers, (2016)

[18] Vadlamannati, K. C. Li, Y. Brazys, S. R. and Dukalskis, A.: Building Bridges or Breaking Bonds? The Belt and Road Initiative and Foreign Aid Competition. SSRN Electronic Journal, (2019)