Financial Engineering and Risk Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2024.070217
ISSN 2523-2576 Vol. 7 Num. 2

Research on the Correlation between Financial News Text Analysis and Stock Market Fluctuations Using Artificial Intelligence Algorithm Models

Yifeng Ye, Rui Ma*

School of Computing, Zhuhai College of Science and Technology, Zhuhai, Guangdong, 519040, China

**Corresponding author*

Keywords: Stock Indices, Financial News, Artificial Intelligence, Model Fundamentals

Abstract: With the vigorous development of the Internet and the rapid popularization of artificial intelligence technology, financial market forecasting and analysis become more convenient and accurate. In this context, this article aims to explore the relationship between financial news and stock price trends, and conduct in-depth analysis using artificial intelligence related algorithms. By utilizing natural language processing techniques, we can quantify and analyse a large amount of financial news, thereby predicting the rise and fall trends of stock indices. We adopted artificial intelligence related technologies, combined with financial news text data, and used a series of algorithm models for analysis. By annotating the rise and fall of stock indices and corresponding news text data, and conducting multiple independent experiments, we divided the dataset into training and testing sets to verify the accuracy and reliability of our model. The research results show that financial news has a certain degree of accuracy in predicting the rise and fall of stock indices, and we observe that there is a certain lag in the response of stock indices to financial news. This means that the stock market's response to news is not immediate, but rather has a certain degree of delay. In addition, we also found some correlation differences between different markets. The research results of this article not only provide new ideas and methods for predicting and analysing financial markets, but also provide a new perspective for people to better understand the relationship between financial news and stock price trends.

1. Introduction

Manufacturing is the cornerstone of the real economy, which in turn acts as the bedrock for China's development and plays a pivotal role in establishing strategic advantages for future growth. At present China's manufacturing scale has been the world's leading position for many years in a row. However, amid the backdrop of the new global manufacturing landscape, China's manufacturing sector still confronts fundamental challenges such as deficiencies in basic research and development and key core technologies, as well as a reliance on imported industrial development.

The high-quality development of the manufacturing industry is the top priority for the high-quality development of China's economy. Innovation serves as the primary catalyst for propelling the high-quality development of the manufacturing industry. Promoting the high-quality development of the industrial chain is a significant foundation for building a new development pattern. The innovation chain is a process from the original idea to product marketization, the industrial chain is the interconnected form of production factors that are shaped based on the upstream and downstream relationships as well as the spatial layout of the production process. It serves as a tangible manifestation of innovative accomplishments. Currently, despite significant advancements in China's scientific and technological innovation capabilities, they still fall short of meeting the demands for high-quality economic development. The persistent issue of a "two-tiered" relationship between scientific and technological progress and economic growth persists [1]. The 20th CPC National Congress Report clearly proposed to "promote the deep integration of innovation chain with industrial chain. Therefore, promoting the organic matching and interactive integration of the innovation chain and the industrial chain has become an inevitable choice for China under the new situation.

Although many scholars have studied the meaning, nature and relationships of the innovation chain and the industrial chain, there is still a gap in the existing knowledge system. Based on the above problems, the following exploration is conducted: First, the impact of the use of fixed effect models to check the innovation chain on the high -quality development of the manufacturing industry. Furthermore, to delve deeper into the analysis of the internal mechanism of integrating innovation, we employ the intermediary effect model to scrutinize the intermediary role of the industrial chain in the relationship between the development of the innovation chain and the high-quality development of the manufacturing industry.

2. Related Research

Social media applications are now the main source of research for many, and also the primary source of information for many users to make decisions based on easily accessible information. S Weshah et al. investigated the impact of social media applications on stock price prediction and trading volume in three categories of the Amman Stock Exchange - ASE (First, Second, and Third Markets)[1]. The results show that social media applications have a significant impact on predicting stock prices and trading volume in the first market, but have a moderate impact on the second and third markets. F Zeng mentioned in the article that with the rapid development of online media, the coverage of online media has more or less influenced the psychological level of investors [2]. Research has found a positive correlation between media attention and investors' heterogeneous beliefs, indicating that investors are more inclined to choose stocks that are frequently reported by the media. Further research has found that media coverage strengthens investors' heterogeneous beliefs, affects their investment behavior, and ultimately leads to an increase in stock trading volume. W Xu et al. believe that stock trend prediction aims to predict the future trend of stocks and is crucial for investors to seek maximum profits from the stock market [3]. Many event driven methods utilize events extracted from news, social media, and discussion forums to predict stock trends in recent years.

CC Wu et al. studied a model of the impact of real negative news on stock prices in this article, and provided evidence using Chinese A-share listed companies as an example [4]. Research has shown that negative news one day and four days later can cause greater stock price volatility, resulting in excess returns. Negative online news can have a certain degree of impact on the stock price fluctuations of listed companies; In the short term, the stock prices of listed companies fluctuate more actively. This article finds that when negative information about listed companies is

disclosed by online media from the day before to the four days after, the stock price will fluctuate significantly, generating excess returns and continuously making stock volatility more active in the short term. B Lei et al. constructed a theoretical model based on the HAR-RV model to analyze the contagion structure between media sentiment and investor sentiment, as well as their impact on stock market performance. The empirical results demonstrate that the level of optimism reported by the media positively affects the subjective emotions of investors and increases their trading volume [5]. Media sentiment can indirectly affect the excess returns and volatility of stocks through investor sentiment and trading.

3. Model Fundamentals and Data Source Processing

3.1 Classification and Regression Tree

Classification regression tree is a commonly used algorithm in machine learning, which can handle classification and regression tasks. The decision tree algorithm includes various variants, with the most common being C4.5, ID3, and CART. The structure of a decision tree is like a tree, with each node representing a feature attribute, and each branch and leaf node representing a possible classification result. The CART algorithm has flexibility and can generate classification or regression trees based on the data type of the predicted results. For classification trees, CART uses the Gini coefficient to evaluate the quality of node splitting. The smaller the Gini coefficient, the better the feature selection effect of the model; for regression trees, CART uses the minimum variance of the sample as the basis for node splitting. Simplify the binary classification problem to the expression of Gini coefficient in the article:

$$Gini(p)=\sum P_k(1-P_k)=2P(1-P) \tag{1}$$

Assuming that a certain type of feature B has two attributes, divide the sample into $|D_1|$ and $|D_2|$, the expression for the Gini coefficient is:

$$Gini(D,B)=\frac{|D_1|}{D}Gini(D_1)+\frac{|D_2|}{D}Gini(D_2) \tag{2}$$

3.2 Support Vector Machines

Support Vector Machine (SVM) is a machine learning model used for binary classification, which is based on the VC dimension theory of statistical learning and the principle of minimizing structural risk. The core idea of this model is to find an optimal balance between the complexity and fitting ability of the model, in order to improve its generalization ability and robustness. It classifies based on the positional relationship of training samples in space. The basic goal is to find a hyperplane that can correctly partition the training data and maximize the geometric spacing between the data. The advantage of this method is that it can effectively process high-dimensional data and in some cases, it can handle non-linear classification problems through kernel techniques.

Based on scientific, systematic and data availability, the evaluation index system of innovation chain is established from three aspects: innovation input, innovation output and innovation value.

3.3 Random Forest

One of the advantages of random forest as an ensemble learning method is its ability to

effectively handle high-dimensional data and large-scale datasets. By randomly selecting features for training, random forests are less susceptible to noise and over fitting, and have good generalization ability. In addition, since each decision tree is trained independently, it can be processed in parallel, improving the speed of model training and prediction, making it particularly suitable for applications in large-scale data scenarios. In addition, random forest can also evaluate the importance of features, analyze the contribution of each feature in the model, help us understand the key features of the data, and guide the process of feature selection and data preprocessing. The evaluation of the importance of this feature is of great significance for a deeper understanding of the patterns and trends behind the data, which helps optimize the model and improve prediction accuracy.

3.4 Model Result Evaluation System

In binary classification tasks, we usually use four performance evaluation metrics to evaluate the performance of the model. These indicators include accuracy, precision, recall, and F1 score. Accuracy measures the overall accuracy of the model's predictions, while precision measures the accuracy of the model's predictions when they are positive. The recall rate measures the model's ability to recognize positive examples. At the same time, the F1 score takes into account both precision and recall, and is the harmonic average of these two factors. Unify the two indicators, with a maximum value of 1 and a minimum value of 0. The calculation formula is as follows:

$$F1-Score=\frac{2\times \Pr ecision\times \operatorname{Re} call}{\Pr ecision+\operatorname{Re} call} \tag{3}$$

3.5 Data Source and Processing

The data source for this article has chosen news text data published online, because online news has many advantages, such as fast information dissemination speed, high timeliness, and convenient access. Especially in the investment field, investors need to timely understand the macroeconomic situation of the country, financial and securities market trends, and financial information of specific stocks corresponding to enterprises. Online news is one of the important channels to provide this information. We used web scraping technology to obtain title data of financial news from websites such as Sina Finance. Cleaning news headlines is an essential step in data processing to ensure the quality and accuracy of the data. Subsequently, word segmentation was performed on the cleaned news headlines in order to transform the text into a data form that can be quantitatively analyzed. The precise mode of the Jieba word segmentation tool can handle Chinese text well, while removing stop words processing helps to eliminate common vocabulary that is useless for analysis, improving the quality and accuracy of the data. We used the TF-IDF model to quantitatively represent the segmented text.

4. Empirical Analysis

4.1 Title Prediction for the Day: Empirical Results of the Rise and Fall of the Shanghai Composite Index

According to the table 1 analysis, the random forest model has achieved astonishing success in accuracy, reaching 64.86%. On the contrary, the worst classification regression tree model had a full 9.91 percentage points lower accuracy. In terms of accuracy, Random Forest also performed outstandingly, reaching 64.83%. This result is 2.98 percentage points higher than the lowest support

vector machine model. However, in terms of recall, support vector machines perform the best. This number is a staggering 31.82 percentage points higher than the worst classification regression tree model. In addition, random forests also showed significant performance in F1 Score and AUC, reaching 75.15% and 64.0%, respectively. These numbers are 14.22 percentage points and 10.0 percentage points higher than the worst classification regression tree model, respectively.

Table 1: Title prediction for the day: Empirical results of the rise and fall of the Shanghai Composite Index

	Support Vector Machine	Multi-layer perceptron	Random Forest	Classification regression tree
Accuracy	0.6126	0.5945	0.6486	0.5495
Precision Rate	0.6185	0.6478	0.6483	0.6290
Recall	0.9091	0.6970	0.8939	0.5909
F1-Score	0.7361	0.6715	0.7515	0.6093
AUC	0.601	0.600	0.640	0.540

It can be seen that this model has high accuracy and reliability in predicting the rise and fall trends of the Shanghai Composite Index. The accuracy and excellent performance of F1 Score mean that the model can make fewer errors in predicting market trends, providing reliable decision-making basis. Meanwhile, although the recall rate of random forest is slightly lower than that of support vector machine, it still reaches a relatively high level, indicating that the model can effectively capture the proportion of real ups and downs, which is crucial for investors to make correct market judgments. By comparing the worst classification regression tree model, we can see significant advantages of random forest in various evaluation indicators, which is attributed to its ability to reduce overfitting and improve the model's generalization ability by integrating multiple decision trees. In addition, the AUC value of the random forest is also relatively high, indicating that the model has a larger area under the ROC curve, indicating that the model has strong ability to distinguish between positive and negative cases. The phenomenon of recall rate being higher than accuracy indicates a closer correlation between the decline of the Shanghai Composite Index and Chinese news on that day. This may reflect that the market tends to be more conservative and cautious in responding to information, that is, paying more attention to factors that may lead to a market downturn.

4.2 Empirical Results of Predicting the Rise and Fall of the Shanghai Composite Index the Next Day Based on the Title of the Day

Table 2: Empirical results of predicting the rise and fall of the Shanghai Composite Index the next day based on the title of the day

	Support Vector Machine	Multi-layer perceptron	Random Forest	Classification regression tree
Accuracy	0.5945	0.6306	0.6667	0.6216
Precision Rate	0.5980	0.6428	0.6522	0.6667
Recall	0.9384	0.8307	0.9231	0.7076
F1-Score	0.7305	0.7248	0.7643	0.6865
AUC	0.624	0.618	0.557	0.573

According to the experimental results is shown in Table 2, the random forest model performed well in predicting the rise and fall trends of the Shanghai Composite Index the next day, with an accuracy of 66.67%, which is 7.22 percentage points higher than the lowest support vector machine model. In terms of recall, the support vector machine achieved 93.84%, which is 23.08 percentage

points higher than the worst classification regression tree model. These numbers show that the Chinese news of the day has considerable reference value for predicting the rise and fall trends of the Shanghai Composite Index the next day. However, the classification regression tree model with the highest accuracy reached 66.67%, slightly higher than the random forest model. This may mean that there is a certain balance and trade-off between accuracy and recall when predicting market trends. Meanwhile, it should be noted that the Dow Jones Index and the Shanghai Composite Index perform similarly in qualitative analysis, both of which are influenced by similar news factors. However, there are certain differences in the quantitative results. This difference can be attributed to the comprehensive influence of various factors, such as investor structure, cultural background, trading system, etc. Especially in the Shanghai Composite Index market, the presence of noise traders may have a significant impact on market reactions, leading to a certain lag in market reactions to negative news. This also means that the actual impact of news information on the market may not be fully synchronized with the time of information release, and multiple factors need to be comprehensively considered for prediction and analysis in order to more accurately understand and predict market trends.

5. Summary and Suggestions

5.1 Summary

Through the research and empirical analysis in this article, we have drawn some key conclusions. Firstly, methods based on natural language processing have shown significant potential and effectiveness in quantifying financial news and stock information. By annotating news texts with stock index prices and finely adjusting the prediction model, we have proven that there is indeed a certain degree of correlation between news texts and stock indices. Secondly, after multiple independent experimental verifications, we have concluded that using relevant news text information to predict the rise and fall trends of the Dow Jones and Shanghai Composite Index is feasible and has significant effects. This discovery further consolidates the correlation between news texts and stock indices, thus supporting our research hypothesis. At the same time, we observed that the impact of news on the stock index of the same day and the next day varies. The rise and fall of the stock index on the next day are more sensitive to the relevant news of the previous day, indicating that there is a certain degree of time lag in the stock market's response to news. Finally, we found that there are differences in the degree of correlation between the rise and fall of stock indices and relevant news. In most cases, the decline of stock indices is more closely related to relevant news. Overall, there is indeed a certain correlation between the text analysis of financial news and stock market volatility. This correlation not only reflects the sensitivity of market participants to news information, but also reflects changes in market sentiment and expectations. Therefore, our research provides useful references and insights for understanding and explaining the complex relationship between financial news and the stock market.

5.2 Suggestions

5.2.1 For Small and Medium-Sized Investors

We should focus on long-term investment strategies and risk management, rather than being influenced by short-term market fluctuations and media news. To establish one's own investment philosophy and methodology, deeply understand the fundamentals and industry trends of investment targets, and avoid blindly following trends and emotional investment decisions. In addition, diversified investment portfolios can be considered to diversify risks and improve investment

returns.

5.2.2 For Listed Companies and News Media Practitioners

It is necessary to strengthen moral and sense of responsibility to avoid damaging market fairness and investor interests for short-term benefits. Listed companies should actively participate in social responsibility, maintain their corporate image and brand value, while news media practitioners should abide by professional ethics in journalism, report financial information impartially, and provide investors with objective, comprehensive, and truthful information.

5.2.3 For Regulatory Agencies

It is necessary to establish a sound information supervision system and investor protection mechanism, strengthen supervision and disposal of market chaos and bad behavior, and maintain the stability and healthy development of the financial market. At the same time, we need to promote the internationalization and transparency of the financial market, attract more high-quality capital to enter the market, and promote a virtuous cycle and long-term development of the capital market.

References

[1] Weshah S, Alazzam E, Aldabbas Q, et al. The Impact of Social Media Applications on Predicting Stocks Prices and Exchange Volume: The Case of Jordan. Allied Business Academies, 2021, 56-69.

[2] Zeng F. Influence of Media Attention on Investors Heterogeneous Beliefs: A Case Study of China Stock Market. Journal of Economic Management: Chinese and English versions, 2021, 10(1):9.

[3] Xu W, Liu W, Xu C, et al. REST: Relational Event-driven Stock Trend Forecasting. Papers, 2021, 89-97. DOI:10.1145/3442381.3450032.

[4] Wu C C, Yan Y, Yuan T, et al. A Study of Network Negative News Based on Behavioral Finance Analysis of Abnormal Fluctuation of Stock Price. Discrete Dynamics in Nature and Society, 2022, 20-22.

[5] Lei B, Song Y. The impact of contagion effects of media reports, investors' sentiment and attention on the stock market based on HAR-RV model. International Journal of Financial Engineering, 2023, 10(02):5-9. DOI:10. 1142/ S242478632350010X.