doi:10.20944/preprints202306.0323.v1

Article

Utilizing Text Mining for Labeling Training Models from Futures Corpus in Generative AI

Hsien-Ming Chou [1,*] and Tsai-Lun Cho [2]

[1] Chung Yuan Christian University; Taiwan
[2] National Tsing Hua University; Taiwan; saetnsaetn@uch.edu.tw
* Correspondence: chou0109@cycu.edu.tw

Abstract: For highly time-constrained very short-term investors, reading and extracting valuable information from financial news poses significant challenges. The wide range of topics covered in these news articles further compounds the difficulties for investors. The diverse content adds complexity and uncertainty to the text, making it arduous for very short-term investors to swiftly and accurately extract valuable insights. Variations in authors, media sources, and cultural backgrounds also introduce additional complexities. Hence, performing a bull-bear semantic analysis of financial news using text mining technologies can alleviate the volume, time, and energy pressures on very short-term investors while enhancing the efficiency and accuracy of their investment decisions. This study proposes labeling bull-bear words from a futures corpus detection method that extracts valuable information from financial news, allowing investors to understand market trends quickly. Generative AI models are trained to provide real-time bull-bear advice, aiding investors in adapting to market changes and devising effective trading strategies. Experimental results show the effectiveness of various models, with Random Forest and SVMs achieving an impressive 80% accuracy rate. MLP and Deep learning models also perform well. By leveraging these models, the study reduces the time spent reading financial articles, enabling faster decision-making and increasing the likelihood of investment success. Future research can explore the application of this method in other domains and enhance model design for improved predictive capabilities and practicality.

Keywords: text mining; semantic analysis; labeling bull-bear words; futures corpus; generative AI

1. Introduction

Currently, the internet is flooded with fake news and futures transactions designed to mislead retail investors. These deceptive practices enable large investors to manipulate news or exploit public information, leading to widespread incorrect judgments among most investors. Generative AI has the potential to train models capable of analyzing and discerning news texts, automatically identifying the distinguishing features of fake news [1]. The model has the ability to learn patterns in the text, identify emotional biases, and recognize the distinguishing factors that separate fake news from genuine information. This knowledge can assist investors in discerning between real news and potential disinformation, enabling them to make more informed decisions [2]. Utilizing a model trained by generative AI, investors can gain valuable insights into the prediction and potential impact of news on bull-bear trends. This guidance empowers investors to make more accurate investment decisions, resulting in improved outcomes [3]. Nevertheless, the labeling process involved in the pre-processing of these models is currently fraught with challenges [2]. The performance and accuracy of generative AI models are inherently tied to the quality and reliability of the training data. If the training data contains errors, biases, or inaccurate information, it can lead to misleading results generated by the models [4]. In the field of finance and economics, labels often adhere to general rules, but in more specific areas such as assessing index futures, the duality of bull and bear positions complicates the labeling process. The overall direction of news may not be immediately apparent, and it may require comprehensive analysis across multiple sources. Additionally, there is currently a shortage of labels that capture the impact of news broadcasts that induce investor panic [5].

The primary focus of this research lies in addressing the challenges faced by current researchers, particularly in regards to the labeling of bull-bear news on semantic analysis for index futures. The ultimate goal is to establish a comprehensive futures corpus that can serve as an effective training resource for generative AI models. Hence, based on this literature review, we have identified two primary research questions in this field. Firstly, what are the essential input factors for training, i.e., what are the key features that machine learning relies on for accurate labeling? Secondly, does the process of labeling contribute positively to the detection of articles by generative AI, thereby enhancing the accuracy rate and facilitating the creation of a profitable training model? To address the aforementioned research questions, this study will commence by conducting a comprehensive literature review to ascertain if existing research methodologies have successfully answered these questions or if there are any limitations encountered by previous researchers. Subsequently, this study will propose innovative research methods aimed at empirical investigation through a research model, with the objective of validating the proposed hypotheses. Drawing from the empirical findings, the study will endeavor to provide explanations, discuss potential discoveries, and ultimately draw conclusions.

The remainder of this paper is organized as follows: The next section provides a review of the related literature. Section 3 details research methods of labeling training model from futures corpus. Sections 4 and 5 present experimental process, results, and discussion. Finally, the last section summarizes our contributions and provides suggestions for future research.

2. Literature Review

There has been a notable emphasis on generative algorithms in recent research endeavors [6-9]. Nevertheless, in more specialized and niche fields, it is often necessary to provide a specific direction or concise summary of the text [10]. In the context of index futures, particularly in the case of relatively short-term texts, making instantaneous judgments becomes challenging. This difficulty arises due to the limited availability of labeled articles in this specific field, resulting in insufficient or incorrect learning materials. When learning is based on inaccurate or inadequate sources, even achieving a high correct rate may lead to erroneous outcomes due to overfitting, which is not beneficial to the majority of investors. For example, Gurrib et al., in their study published in 2022 [11], concentrated on commodity futures and did not undertake a labeling process specifically for index futures. Similarly, Dai et al. in their study published in 2022 [12], and Wang et al. in their study published in 2020 [13], did not specifically address the labeling process for index futures. In these recent studies, addressing the first research question regarding the input factors for training and labeling, providing a comprehensive explanation, especially for index futures with short-term fluctuations, is challenging. Due to the dynamic nature of index futures and the varied influencing factors at play, it becomes difficult to identify a specific set of input factors that consistently apply across different scenarios. The uniqueness and complexity of short-term fluctuations in index futures result in significant variations in the influencing factors [14, 15].

Indeed, prior to commencing training of a generative AI model, a substantial amount of preparatory work needs to be undertaken in the text domain. This includes tasks such as data collection, where text data is gathered to serve as training material for the model [15, 16]. Accurately stated. In order to collect the necessary text data for training, various methods can be employed, such as utilizing web crawlers or executing targeted database queries. Once the data is collected, the next crucial step is data cleaning and preprocessing. This involves removing special characters, punctuation marks, numbers, and other unnecessary elements from the collected data. Additionally, handling missing or erroneous data is also part of the preprocessing phase to ensure the data is of high quality and suitable for training the generative AI model [17]. Tokenization or sentence segmentation involves splitting text into sequences of words or sentences. This process can be achieved by utilizing simple methods such as whitespace or punctuation-based tokenization, or more sophisticated techniques using natural language processing for advanced tokenization or sentence segmentation [18]. To create a vocabulary, one must build it based on the word segmentation results, encompassing all the words or phrases present in the training data. The vocabulary acts as a comprehensive collection of unique terms derived from the training data [18, 19]. Data encoding is a crucial step in converting text data into a numerical representation that can be effectively processed by the model. Several common approaches are employed for this purpose, such as one-hot encoding, word embedding, or other vectorization techniques. These techniques enable the transformation of textual information into numerical values, facilitating the model's ability to work with the data [18].

In order to train text data, it is necessary to build a suitable model and select an appropriate architecture. There are several options to consider, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), or transformer models, among others. Each architecture offers unique advantages and is applicable in different scenarios. The choice of model architecture should be based on the specific characteristics of the text data and the objectives of the task at hand [20]. Once the selected model has been established, the next step is to train it using the training data. This process involves feeding the data into the model for forward propagation, calculating a loss function to measure the model's performance, and utilizing the backpropagation algorithm to update the model's parameters. Additionally, parameter adjustment is essential, where model parameters are fine-tuned based on the training process results and their impact on performance.

This adjustment may involve tweaking hyper parameters like learning rate and regularization to optimize the model's performance [21]. Determining whether the annotation process is beneficial for article detection by generative AI, leading to improved accuracy and the development of a profitable training model, requires validation and testing. Trained models need to be evaluated using separate validation datasets to assess potential overfitting and overall performance. This process is complex and essential to answer the second research question. However, existing studies lack focus on index futures, particularly in the context of extremely short-term bull-bear texts, which makes it challenging to generate a fully labeled training set specifically for this domain [22]. To address the need for a comprehensive semantic library or corpus for subsequent generative AI training, it is essential to either compile an existing semantic library or create a new corpus specifically tailored for the desired training objectives. This involves gathering relevant and diverse text data from reliable sources, ensuring the inclusion of various linguistic patterns and domain-specific knowledge. By curating a robust semantic library or creating a new corpus, researchers can provide valuable resources for training generative AI models effectively.

3. Labeling Methods for Generative AI

Based on the research methods and issues discussed in recent literature, this study employs text processing techniques aligned with generative AI to collect text data and perform various preprocessing steps for the domain of extremely short-term trading in index futures. The methodology involves creating and labeling a bull-bear training set, as depicted in Figure 1. Initially, a widely recognized and extensive financial website is utilized as a web crawler to gather relevant articles. When employing deep learning for natural language processing, the articles are transformed, converting each word into a vector representation through techniques such as word embedding.

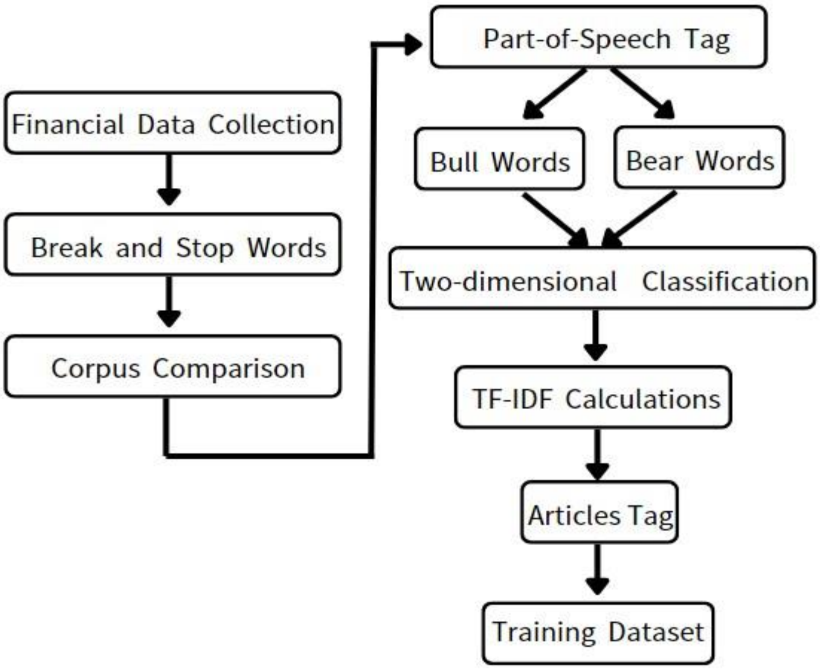

Figure 1. Bull and Bear Words Training Set.

In Figure 1, pertained word vectors can be employed to convert words into vectors. One widely used approach is Word2Vec, which is a neural network-based language model that represents words as high-dimensional vectors. In the context of Chinese financial news, Word2Vec can be utilized to train word vectors by considering multiple words or even word sequences as a single unit. This approach, known as a word vector model using multiple words or word sequences as a unit, enhances the representation power of the word vectors and captures more contextual information. Chinese words are often structurally complex, comprising multiple characters or morphemes. Consequently, Chinese text processing commonly involves word segmentation to divide the text into individual words or phrases. In the Word2Vec model, the segmented words serve as training units to derive vector representations for each word. These word vectors effectively capture semantic information and express contextual relationships between words in Chinese. Subsequently, the vector representations of all words are concatenated to create a matrix of dimensions (N, d), where N represents the number of words in the text and d represents the dimension of the word vector. This matrix can be perceived as a sequence of length N. Moving forward, the sequence is subjected to a 1D convolutional layer. The convolution operation is applied to the sequence, allowing flexibility in defining the size of each convolution kernel.

The convolution operation treats neighboring words as local text segments, enabling the extraction of features from these segments. It can be visualized as a sliding window moving across the sequence, extracting features within the window at each step. Subsequently, a pooling layer is applied to reduce the extracted features from each convolution kernel to a single numerical value. Common pooling techniques, such as Max Pooling or Average Pooling, can be utilized for this purpose. This step aids in compressing the features extracted by each convolution kernel, resulting in a reduction in the number of model parameters. Finally, all the compressed features are concatenated and passed through a fully connected layer for classification. In this specific case, a three-layer fully connected network can be employed for classification. The number of neurons in the first layer can be set to match the feature dimension outputted by the pooling layer, while the number of neurons in the second layer can be determined by the number of categories for text classification. The "Softmax" activation function is commonly used for classification tasks. In summary, the proposed deep learning text classification model utilizes convolution operations to extract sequence features and employs pooling layers for feature compression. This allows the text to be represented as a fixed-dimensional vector suitable for subsequent classification tasks. The model design can be adapted for various text classification applications and can be further optimized by adjusting hyper parameters such as the size, number, and pooling methods of the convolution kernels. Additionally, part-of-speech tagging, specifically Chinese part-of-speech tagging, involves assigning each word in a Chinese sentence its corresponding part of speech, such as noun, verb, adjective, etc. For unknown words, their part of speech can be inferred from the dictionary entry for that word.

Through the analysis of sentence structure, word semantics, and linguistic rules, part-of-speech tagging can be performed by assigning each word its appropriate part of speech. In addition, TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used text mining technique that can be applied to evaluate the importance of each word within the financial news corpus. It calculates the word frequency of each term within the document and then determines the inverse document frequency of each term using the IDF formula. By multiplying the TF and IDF values, the TF-IDF value of each word can be obtained. By calculating the TF-IDF values for both bull and bear words, the final classification of the article as either bull or bear can be determined. This process involves collecting a significant number of financial articles, annotating them accordingly, and assembling them into a training set for the subsequent steps.

The second step involves constructing a futures corpus, as depicted in Figure 2. Initially, financial articles need to be collected and compiled into a text dataset. These articles consist of news reports sourced from reputable financial websites. The relevance of these texts to bull and bear futures is verified through the reports provided by financial experts on the websites. Once the labeled training set is available, the entire text dataset is utilized as a corpus. Pre-processing techniques, such as synonym and antonym identification, are employed to establish the bull and bear terms. Additionally, various text exploration techniques are applied to the generated training set, including tasks like word segmentation, word frequency statistics, and TF-IDF calculations to determine the weights of vocabulary. Through processing and analysis of the corpus, more valuable text information can be obtained.

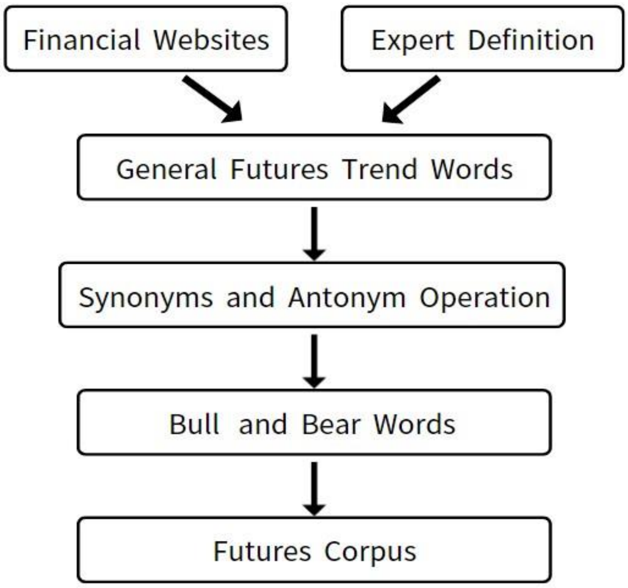

Figure 2. The Futures Corpus.

Once the corpus is established, a portion of the text can be chosen as a test set for model evaluation and validation purposes. Furthermore, the training set can be retained within the corpus to facilitate the retrieval of suitable training data for subsequent model training. In summary, it is a common practice to first construct a training set and then build a corpus. This approach allows for a more targeted construction of the corpus, enhancing the efficiency and accuracy of text mining tasks. By ensuring a focused and relevant training set, the subsequent corpus can be designed with greater precision, leading to improved outcomes in text analysis and mining processes. The third step involves making real-time predictions on articles that can influence actual bull-bear trends, as illustrated in Figure 3. This part focuses on capturing relevant real-time financial news, inputting it into the trained training set, and retrieving the appropriate training data from the established corpus based on the article's type. Instant predictions can then be generated. Various algorithms can be employed for this purpose, with deep learning being a notable approach that can be considered.

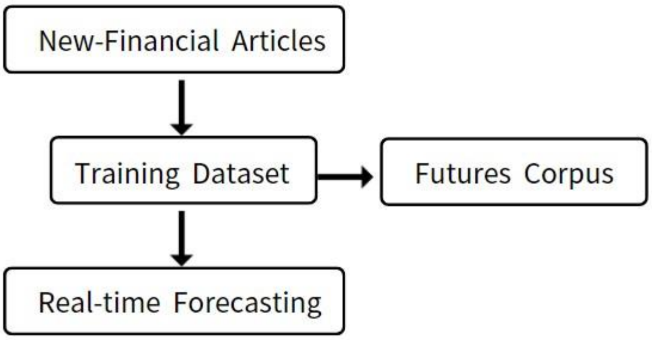

Figure 3. The Effect of Real-time Forecasting.

In summary, the process of capturing and analyzing real-time financial news to make instant predictions on the bull-bear trends of futures involves several key steps. Firstly, relevant financial news needs to be captured from multiple sources using crawlers or similar methods, followed by performing text analysis on the collected news. Secondly, a corpus must be established to store classified futures bull-bear financial news, serving as a training data source for the prediction model. Thirdly, the model needs to be trained using techniques like deep learning, enabling it to automatically analyze news and predict their impact on actual trends, such as bull and bear futures. Real-time financial news is then input into the trained model to obtain predictions on the actual bull or bear positions and their influence on factors like futures prices. It is important to acknowledge the challenges and limitations involved in this process. The vast amount of news information requires quick and accurate analysis. Additionally, the factors influencing actual bull-bear trends, such as futures prices, are complex and require comprehensive consideration during correlation analysis. Therefore, careful model design and training are necessary to address these challenges effectively.

Before proposing the experimental process, a research model is constructed, as shown in Figure 4. The left side represents the pre-processing part of text mining. Various calculations, including PoS Tagging and TF-IDF, are performed to generate multiple empty words. These empty words serve as the training set, with some extracted from the corpus to save collection time. After feedback, the valid data is saved back to the corpus for subsequent extractions. To evaluate the effectiveness of the training models, the latest real-time articles from actual financial websites are obtained for prediction and judgment. These real-time data will be labeled to increase the effectiveness of the training set. The effectiveness of the model needs to be assessed through training and testing using various algorithms. To cover a wide range of possibilities, several popular algorithms such as Support Vector Machines (SVMs), Multi-Layer Perceptron (MLP), Random Forest, and Deep Learning will be used for efficiency verification.

doi:10.20944/preprints202306.0323.v1

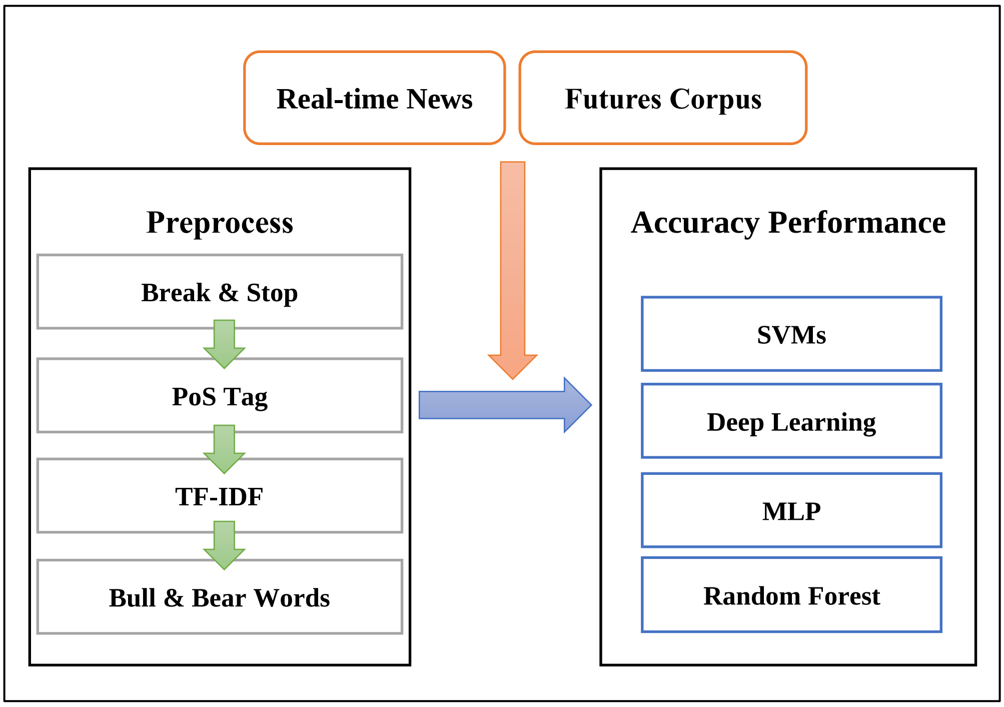

Figure 4. The research model.

4. Evaluation

For the research study, we collected financial news articles from well-known websites such as Juheng.com (https://news.cnyes.com/news/id/5159323) and Business Times (https://ctee.com.tw/news/futures/852930.html) from January 2023 to May 2023. A total of 50 news articles were collected during this period. These articles were subjected to text exploration and processing to extract relevant information. Based on the analysis, we established a set of 100 bull and bear words (https://drive.google.com/file/d/19SqTt-MMYiG4yjxPALbCTb9ECRYebu7SM/view?usp=sharing) that are significant for futures trading. Using the established corpus, we conducted article training to train the model. The training process involved utilizing various techniques such as word segmentation, part-of-speech tagging, and TF-IDF calculations to process the text data. By employing these methods, we aimed to enhance the accuracy and effectiveness of the training model. It is important to note that this research study focused specifically on the domain of futures financial news, and the collected articles were from a specific timeframe. This allowed us to gather a targeted dataset and ensure its relevance to the research objectives.

In machine learning, various metrics are used to evaluate the performance of a model. Some commonly used metrics include accuracy, precision, and recall, which provide insights into different aspects of the model's performance. Accuracy is the proportion of correct predictions among all predictions made by the model. It measures the overall prediction accuracy of the model. A higher accuracy indicates a stronger predictive ability of the model. Precision is the proportion of true positive predictions among all positive predictions made by the model. It measures the model's ability to correctly predict positive examples. A high precision indicates that the model is accurate in identifying positive examples. Recall is the proportion of true positive predictions among all actual positive examples in the dataset. It measures the model's ability to detect positive examples. A high recall indicates that the model can effectively capture positive examples. These metrics are important for evaluating the performance of a model, as they provide insights into its predictive accuracy, ability to identify positive examples, and ability to capture all relevant positive examples. Depending on the specific objectives and requirements of the research or application, different metrics may be given more importance.

To put it simply, the accuracy rate assesses the overall predictive capability of the model, while the precision rate and recall rate evaluate the model's predictive capability and detection ability for positive examples, respectively. Typically, there exists a trade-off between precision and recall, meaning that as we increase precision, recall tends to decrease, and vice versa. In practical applications, we select different metrics based on the specific problem at hand. For instance, in a medical diagnosis problem, our focus is on detecting actual patients, which emphasizes recall. On the other hand, in a credit card fraud problem, our priority is detecting fraud cases, placing greater importance on precision. When dealing with class imbalance, where there is a significant difference in the number of positive and negative examples, relying solely on accuracy may yield misleading outcomes. For instance, let's consider the prediction of tumor patients. If only 1% of the samples in the training set are actual tumor patients, a model that simply predicts all samples as non-tumor would still achieve 99% accuracy. However, such a model would not be useful in practice as it fails to detect any actual tumor cases. This highlights the limitation of relying solely on accuracy when dealing with imbalanced datasets. On the other hand, when dealing with balanced classes where the number of positive and negative examples is equal, using accuracy can be a better choice. For instance, when predicting the bull-bear trend of the stock market, where the probability of bull or bear is approximately 50:50, accuracy is a more suitable metric. It accurately reflects the model's performance in predicting the correct bull-bear trend, taking into account both true positives and true negatives.

The selection of an appropriate evaluation metric depends on the class balance of the dataset. When there is a balanced distribution of positive and negative examples, the accuracy rate can better assess the model's performance. However, when there is a significant class imbalance, relying solely on accuracy may yield misleading results. In such cases, alternative metrics like precision, recall, or F1 score should be considered. In the specific experiment, the Random Forest model achieved an accuracy rate of 80%, SVMs also achieved 80% accuracy, MLP achieved 76% accuracy, and Deep learning achieved 66% accuracy (Table 1). These results provide an initial indication of the models' performance, but further analysis and comparison of other metrics are necessary to make a comprehensive assessment.

Table 1. The Accuracy of Algorithms.

Algorithms	Random Forest	SVMs	MLP	Deep learning
Accuracy	80%	80%	76%	66%

5. Discussion

The results suggest that Random Forest and SVMs have performed relatively well, while Deep learning has shown poorer performance. There are several potential reasons for this disparity. One possible cause is the small sample size of the dataset. When the dataset is small, models can easily overfitting, meaning they become too specialized in capturing the idiosyncrasies of the training data and fail to generalize well to unseen data. This can lead to significant differences in performance between different models. To address this issue, one option is to increase the size of the training dataset. Collecting more data can help improve the model's ability to learn patterns and make more accurate predictions. Alternatively, techniques such as cross-validation can be employed to mitigate overfitting and assess the model's generalization performance more effectively. By addressing the small sample size issue and employing appropriate techniques for model evaluation, it is possible to obtain more reliable and meaningful results that accurately reflect the performance of different models.

The quality of the data and the selection of relevant features can significantly impact the performance of a model. Poor data quality, such as missing values or incorrect labels, can lead to inaccurate predictions. Similarly, inadequate feature selection may include irrelevant or redundant features, hindering the model's ability to learn meaningful patterns.

To address these issues, it is crucial to carefully evaluate the data source and its quality. Cleaning the dataset by handling missing values, outliers, or inconsistencies can greatly improve the model's performance. Additionally, employing effective feature selection techniques, such as statistical analysis or domain knowledge, can help identify the most informative features for the prediction task. Moreover, the model's hyper parameters, such as the maximum depth of decision trees or regularization coefficients in SVMs, play a vital role in performance. Optimal hyper parameter selection is crucial for achieving the best results. Techniques like grid search or random search can be used to systematically explore different hyper parameter combinations and identify the ones that yield the highest performance based on evaluation metrics. By addressing data quality issues, refining feature selection, and fine-tuning the model's hyper parameters, the performance of the models can be enhanced, resulting in more accurate predictions.

In summary, achieving optimal model performance requires careful consideration and adjustment of multiple factors. The quality and quantity of data, feature selection, hyper parameters, and inherent assumptions of the model all contribute to its performance. By conducting thorough analysis and making appropriate adjustments to these factors, the model can be optimized to achieve the best possible performance. This iterative process of evaluation and refinement is crucial in order to ensure accurate and reliable predictions.

6. Conclusion

Financial articles contain complex and diverse information, covering economic indicators, policies, regulations, and company operations, among others. As this information is closely tied to market dynamics, investors need to efficiently acquire, analyze, and comprehend it to make informed investment choices. Particularly for short-term investors, who face the challenge of processing large volumes of data within tight timeframes, the ability to swiftly determine whether an article indicates a bull or bear position holds paramount importance for effective investment decision-making.

The bull-bear futures text detection method proposed in this study has made significant contributions by enabling the extraction of valuable information from a vast corpus of financial articles, enabling investors to gain rapid insights into market trends. By training machine learning models, the study has achieved predictive capabilities, providing real-time bull-bear advice to investors and enabling them to promptly adapt to market changes and devise effective trading strategies. Experimental results demonstrated the effectiveness of various machine learning models, particularly Random Forest and SVMs, achieving an impressive accuracy rate of 80%. MLP and Deep learning models also showcased commendable performance. Through the utilization of these models, the study successfully reduced the time investors spend reading extensive financial articles, enabling faster decision-making and increasing the likelihood of investment success. Future research can expand the application of this approach to other domains and enhance model design to improve predictive capabilities and practicality.

Funding: National Science and Technology Council of Taiwan, under the grant number 111-2410-H-033-010.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to no significant risks through the questionnaire.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available in https://drive.google.com/file/d/19SqTtMMYiG4yjxPALbCTb9ECRYebu7SM/view?usp=sharing (accessed on 1 June 2023).

Acknowledgments: This research was funded by the National Science and Technology Council of Taiwan, under the grant number 111-2410-H-033-010.

doi:10.20944/preprints202306.0323.v1

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferhat Hamida, Z., A. Refoufi, and A. Drif, *Fake news detection methods: A survey and new perspectives.* Advanced Intelligent Systems for Sustainable Development (AI2SD'2020) Volume 2, 2022: p. 123-141.
2. Longoni, C., A. Fradkin, L. Cian, and G. Pennycook. *News from generative artificial intelligence is believed less*. in *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.
3. Matsubara, T., R. Akita, and K. Uehara, *Stock price prediction by deep neural generative model of news articles.* IEICE TRANSACTIONS on Information and Systems, 2018. **101**(4): p. 901-908.
4. He, W., Y. Dai, Y. Zheng, Y. Wu, Z. Cao, D. Liu, P. Jiang, M. Yang, F. Huang, and L. Si. *Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022.
5. Ahnve, F., K. Fantenberg, G. Svensson, and D. Hardt. *Predicting stock price movements with text data using labeling based on financial theory*. in *2020 IEEE International Conference on Big Data (Big Data)*. 2020. IEEE.
6. Gao, W. and C. Su, *Analysis on block chain financial transaction under artificial neural network of deep learning.* Journal of Computational and Applied Mathematics, 2020. **380**: p. 112991.
7. Ponomarev, E., I.V. Oseledets, and A. Cichocki, *Using reinforcement learning in the algorithmic trading problem.* Journal of Communications Technology and Electronics, 2019. **64**: p. 1450-1457.
8. Xie, M., H. Li, and Y. Zhao, *Blockchain financial investment based on deep learning network algorithm.* Journal of Computational and Applied Mathematics, 2020. **372**: p. 112723.
9. Kumbure, M.M., C. Lohrmann, P. Luukka, and J. Porras, *Machine learning techniques and data for stock market forecasting: A literature review.* Expert Systems with Applications, 2022: p. 116659.
10. Gharib, C., S. Mefteh-Wali, V. Serret, and S.B. Jabeur, *Impact of COVID-19 pandemic on crude oil prices: Evidence from Econophysics approach.* Resources Policy, 2021. **74**: p. 102392.
11. Gurrib, I. and F. Kamalov, *Predicting bitcoin price movements using sentiment analysis: a machine learning approach.* Studies in Economics and Finance, 2022. **39**(3): p. 347-364.
12. Dai, Z., J. Zhu, and X. Zhang, *Time-frequency connectedness and cross-quantile dependence between crude oil, Chinese commodity market, stock market and investor sentiment.* Energy Economics, 2022. **114**: p. 106226.
13. Wang, L., F. Ma, T. Niu, and C. Liang, *The importance of extreme shock: Examining the effect of investor sentiment on the crude oil futures market.* Energy Economics, 2021. **99**: p. 105319.
14. Chun, J., J. Ahn, Y. Kim, and S. Lee, *Using deep learning to develop a stock price prediction model based on individual investor emotions.* Journal of Behavioral Finance, 2021. **22**(4): p. 480-489.
15. Li, J., G. Li, M. Liu, X. Zhu, and L. Wei, *A novel text-based framework for forecasting agricultural futures using massive online news headlines.* International Journal of Forecasting, 2022. **38**(1): p. 35-50.
16. Doh, T., D. Song, and S.-K. Yang, *Deciphering federal reserve communication via text analysis of alternative fomc statements.* Federal Reserve Bank of Kansas City Working Paper Forthcoming, 2022.
17. Chou, H.-M., *A smart-mutual decentralized system for long-term care.* Applied Sciences, 2022. **12**(7): p. 3664.
18. Hung, C., W.-R. Wu, and H.-M. Chou, *Improvement of sentiment analysis via re-evaluation of objective words in SenticNet for hotel reviews.* Language Resources and Evaluation, 2021. **55**: p. 585-595.
19. Chung, Y.-C., H.-M. Chou, C.-N. Hung, and C. Hung, *Using textual and economic features to predict the RMB exchange rate.* Adv. Manag. Appl. Econ, 2021. **11**: p. 139-158.
20. Chou, H.-M., *A collaborative framework with artificial intelligence for long-term care.* IEEE Access, 2020. **8**: p. 43657-43664.
21. Chou, H.-M., S.-M. Pi, and T.-L. Cho, *An Intelligent Healthcare System for Residential Aged Care during the COVID-19 Pandemic.* Applied Sciences, 2022. **12**(22): p. 11847.
22. Chou, H.-M. and C. Hung, *Multiple strategies for trading short-term stock index futures based on visual trend bands.* Multimedia Tools and Applications, 2021. **80**: p. 35481-35494.