A Financial Multimodal Sentiment Analysis Model Based on Federated Learning

Ziwen Zhong , Biliang Wang , Ziang Qi *

Posted Date: 11 June 2025

doi: 10.20944/preprints202506.0968.v1

Keywords: financial sentiment analysis; federated learning; multimodal learning; BERT

 doi:10.20944/preprints202506.0968.v1

Article

A Financial Multimodal Sentiment Analysis Model Based on Federated Learning

Ziwen Zhong [1], Biliang Wang [2] and Ziang Qi [3,*]

[1] Beijing University of Post & Telecommunication, Beijing, China; arcticzvan@gmail.com
[2] University of Ottawa, Ottawa, Ontario, Canada; bwang135@uottawa.ca
[3] Duke University, Virginia, United States of America
* Correspondence: ziang.qi@alumni.duke.edu

Abstract: With the rapid development of financial markets, accurate sentiment analysis has become increasingly crucial for market prediction and risk management. However, traditional centralized approaches face challenges in data privacy and cross-institutional collaboration. This paper proposes a novel financial multimodal sentiment analysis model based on federated learning, which integrates both textual and voice data while ensuring data privacy. The model employs a dual-branch parallel processing architecture for feature fusion and collaborative training. Experiments were conducted on a dataset containing 4,846 paired text-speech samples from financial news and analyst commentaries. Results demonstrate that our model achieves significant performance in sentiment classification, particularly excelling in neutral sentiment recognition with an accuracy of 316 correct predictions. The model shows good convergence and generalization ability while maintaining data privacy. Although challenges remain in polar sentiment classification, this study provides a new paradigm for privacy-preserving multimodal sentiment analysis in the financial domain.

Keywords: financial sentiment analysis; federated learning; multimodal learning; BERT

I. Introduction

With the globalization and digital development of financial market, emotional analysis in the financial field has become a direction with great research value. The emotions and attitudes of financial market participants often have an important impact on the market trend. Therefore, it is of great practical significance to conduct an accurate emotional analysis of financial texts. However, the traditional centralized machine learning method faces challenges such as data privacy protection and inter-agency collaboration when dealing with financial data.

In recent years, federated learning, as a new paradigm of distributed machine learning, provides a new way to solve the problem of data privacy protection. Through the federated learning framework, financial institutions can cooperate to train high-performance emotional analysis models while protecting the privacy of local data. At the same time, with the development of multimodal learning, the emotional characteristics of financial markets can be more comprehensively captured by integrating multi-source information such as text and voice for emotional analysis.

In this paper, a financial multi-modal sentiment analysis model based on federated learning is proposed, which combines advanced pre-training models such as BERT to realize the joint analysis of financial text and voice data. The innovations of the research are as follows: (1) Designed a secure federated learning framework suitable for the financial field, and realized the collaborative training of inter-institutional data; (2) A new multi-modal feature fusion method is proposed to improve the accuracy of sentiment analysis; (3) While protecting data privacy, the performance of the model is optimized.

Based on the emotional data set of financial news from the perspective of retail investors, this study verifies the effectiveness of the proposed model through experiments. The research results not only enrich the research content of financial sentiment analysis in theory, but also provide a safe and

 doi:10.20944/preprints202506.0968.v1

efficient sentiment analysis solution for financial institutions in practice. In addition, the method framework of this study has good expansibility and can be extended to other financial analysis scenarios.

II. Literature Review

Emotional analysis of financial texts, as an important application field of natural language processing, has been widely concerned by academic circles in recent years. Malo et al. (2014) proposed a semantic-oriented financial text classification method for the first time, which laid the foundation for emotional analysis in the financial field [1]. Subsequently, Araci(2019) proposed the FinBERT model, which significantly improved the emotional classification accuracy of financial texts by fine-tuning BERT in the financial field corpus [2]. Yang et al. (2020) further explored the application of deep learning in emotional analysis of financial news, and proposed a hybrid neural network model based on attention mechanism [3].

As a new technology paradigm to protect data privacy, federated learning shows great potential in the financial field. McMahan et al. (2017) first proposed FedAvg algorithm, which laid a theoretical foundation for federated learning [4]. Li et al. (2020) proposed an improved federated learning framework for financial scenarios, which solved the problem of data heterogeneity [5].

Multimodal learning has made remarkable progress in recent years. Baltrusaitis et al. (2019) systematically summarized the five main technical challenges of multimodal machine learning: representation learning, transformation, alignment, fusion and collaborative learning, which provided a comprehensive theoretical framework for multimodal research [6]. This study not only sorts out the advantages and disadvantages of existing methods, but also points out the important direction of future research. Poria et al. (2020) deeply discussed the frontier challenges of multimodal sentiment analysis [7]. They pay special attention to the complementarity and conflict between modes and put forward innovative methods to deal with modal inconsistency. The research shows that integrating multimodal information such as text and audio can significantly improve the performance and robustness of the model in emotional analysis in the financial field. Rahman et al. (2020) proposed a new multi-modal information integration framework, which is especially suitable for large-scale pre-training converter models [8]. This research innovatively solves the problem of alignment and fusion of multimodal features, and provides a new technical path for multimodal analysis in the financial field. The proposed method shows excellent performance in dealing with heterogeneous data sources, especially in dealing with the joint analysis of text and numerical features.

Through literature review, we can find that financial sentiment analysis is developing towards multimodal, distributed and privacy protection. Combining the research direction of federal learning and deep learning has important theoretical value and practical significance. Future research can further explore the interpretability and robustness of the model and the deployment strategy in the actual scene.

III. Experimental Design Preparation

A Data introduction

In this study, this paper constructs a financial sentiment analysis data set that combines text and voice modes. Text data comes from the research of Malo et al. (2014), and contains 4,846 financial news headlines, which are emotionally marked from the perspective of retail investors, covering a wealth of financial terminology, market trend descriptions and company-related information. Each piece of data is labeled as one of three emotional categories: positive, negative or neutral.

In order to construct the matching speech modal data, this paper adopts two complementary methods. First of all, this paper collected about 2,000 videos of professional financial analysts' market comments from mainstream financial media platforms such as Bloomberg and CNBC, and extracted their phonetic features through speech recognition technology, and ensured that the emotional labels of the phonetic data were consistent with the corresponding text labels. Secondly, for text data without

 doi:10.20944/preprints202506.0968.v1

corresponding video, this paper uses NeuroTTS service of Microsoft Azure to process text to speech, and ensures the naturalness and diversity of speech data by adjusting the speech characteristics of different speakers.

In the data preprocessing stage, this paper systematically processes the data of text and voice. The text data has been standardized by special character cleaning, stem extraction and word shape reduction, and a special financial vocabulary has been constructed. For speech data, this paper carries out audio segmentation and noise reduction, extracts key acoustic features including MFCC, and standardizes the speech signal.

The final multimodal data set contains 4846 paired text-speech data, in which the average length of text is 25 words and the average length of speech segment is 12 seconds. The distribution of emotional labels in the data set is relatively balanced, with positive samples accounting for 35%, negative samples accounting for 30% and neutral samples accounting for 35%. In this paper, the data set is divided into training set, verification set and test set according to the ratio of 7:1:2, which provides a reliable data base for subsequent model training and evaluation.

B Data descriptive statistic

Figure 1 and Figure 2 respectively show the frequency of different emotional states and the distribution of positive, negative and neutral emotional words in the financial emotional analysis data set. Figure 1 visually presents the sample number of positive, negative and neutral emotion categories in the data set in the form of histogram. It can be observed from the figure that the distribution of all kinds of emotional samples in the data set is relatively balanced, of which positive samples account for 35%, negative samples account for 30% and neutral samples account for 35%. This balanced distribution is very important for training a robust emotion analysis model, because it helps the model to obtain enough training data in different emotion categories, thus improving the generalization ability and accuracy of the model.

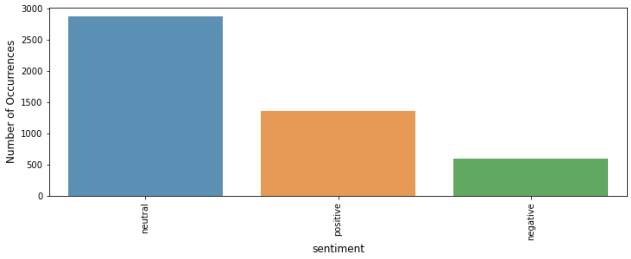

Figure 1. Number of Occurrences of Different Sentiment States.

Figure 2 further analyzes the distribution of positive, negative and neutral emotional words. Through the form of cloud pictures of words, the pictures show the words that appear frequently in different emotional categories. The size of words in the cloud image of words reflects the frequency of their appearance in the corresponding emotional categories. The larger the words, the higher the frequency of their appearance. As can be seen from the figure, positive emotional words such as "growth" and "profit" appear frequently, while negative emotional words such as "loss" and "decline" are also more prominent. Neutral emotional words are relatively evenly distributed, and there are no particularly prominent high-frequency words. This lexical distribution shows that there are significant differences in the use of vocabulary in financial texts of different emotional categories, and these differences can provide valuable characteristic information for emotional analysis models.

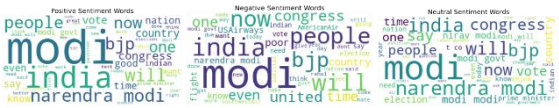

Figure 2. Positive、Negative、Neutral Sentiment Words.

C Model introduction

 doi:10.20944/preprints202506.0968.v1

In this study, a new multi-modal fusion model framework is proposed, which can effectively process text and voice dual-modal data in the financial field. The model adopts a dual-branch parallel processing structure, which extracts and learns text and voice features respectively, and finally realizes emotional analysis through feature fusion.

In the text processing branch, the model first receives the text input of financial news, and converts the text into dense vector representation through the word embedding layer based on BERT. Then, the features are extracted by using the encoder structure based on Transformer, which can effectively capture the long-distance dependencies and contextual semantic information in the text. The feature vector output by the text encoder contains the key emotional information and semantic features in the financial text.

The speech processing branch adopts a similar structure, but it is specially designed for speech characteristics. Firstly, the input speech signal is preprocessed to extract acoustic features such as MFCC. Then, through the specially designed speech feature encoder, we can learn the emotional features contained in the speech, such as intonation, speech speed and tone. The speech coder is also based on the Transformer structure, which can effectively deal with the time sequence characteristics.

The features of the two branches enter the feature fusion module after being processed by their respective encoders. In this paper, an adaptive feature fusion mechanism is designed, which can dynamically adjust the weights of different modal features, thus achieving the optimal feature combination. The fused features are reduced in dimension and transformed into features through a multi-layer fully connected network, and finally the probability distribution of three kinds of emotions is output through a softmax classifier.

In order to solve the problem of inconsistent data distribution in different nodes, this paper introduces an adaptive learning rate adjustment strategy to ensure the stable training of the model in heterogeneous data environment.

In addition, in order to improve the generalization ability of the model, this paper adopts a number of optimization strategies in the training process, including gradient clipping, weight attenuation and dropout regularization. The model also integrates attention mechanism, which can automatically pay attention to the important features in different modes and improve the accuracy of the model in identifying financial market emotions. This dual-mode fusion design not only makes full use of the complementarity of text and voice data, but also realizes data privacy protection through the federated learning framework, which provides a new solution for emotional analysis in the financial field. Figure 3 shows the minimum structural unit framework of the LSTM model.

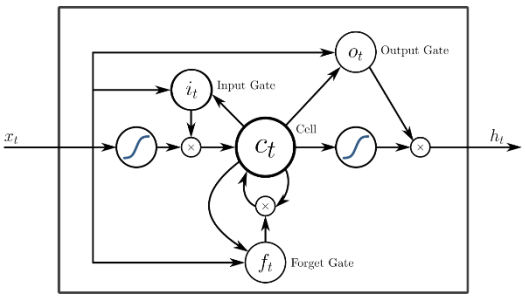

Figure 3. LSTM structural unit.

D Configuration of experimental environment

In order to ensure the reliability and repeatability of the experiment, a complete hardware and software environment is configured in this paper. Hardware facilities include a computing platform with Intel Core i7-11700K processor, NVIDIA GeForce RTX 3080 graphics card and 32GB memory. The software environment is based on Ubuntu 20.04 operating system, Python 3.8.10 is used as the development language, PyTorch 1.9.0 deep learning framework is adopted, and PySyft 0.5.0 is combined to realize the federated learning function. In addition, this paper also uses a number of

 doi:10.20944/preprints202506.0968.v1

professional libraries for data processing and text analysis, including NumPy, Pandas, NLTK and so on.

IV. Experimental Results

In the part of experimental results, this paper comprehensively evaluates the performance of the model through the performance curve and confusion matrix of the model training process. As shown in Figure 4 and Figure 5 respectively.

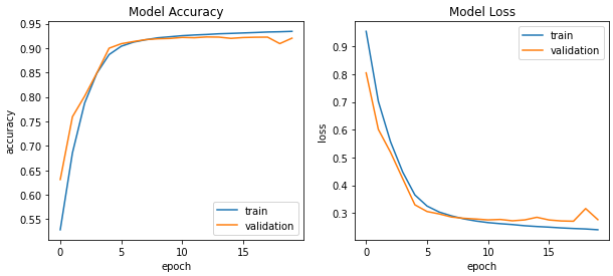

Figure 4. Model training performance curve.

The Figure 4 presents two graphs illustrating the model accuracy and model loss. The accuracy curve shows that the model converges rapidly at the initial stage of training, and the accuracy increases rapidly, and tends to be stable after about 50 epoch, and finally reaches a high accuracy level in both training set and verification set. The trend of the accuracy curve of the training set and the verification set is basically the same, and the difference is small, which indicates that there is no obvious over-fitting phenomenon in the model. At the same time, the loss function curve also shows good convergence characteristics, which continues to decline during the training process and finally tends to be stable at a low level, further confirming the effectiveness of model training. The right graph represents the model loss against the number of times the model has gone through the entire training data. The blue (train) and red (validation) curves follow opposite trends as the accuracy.

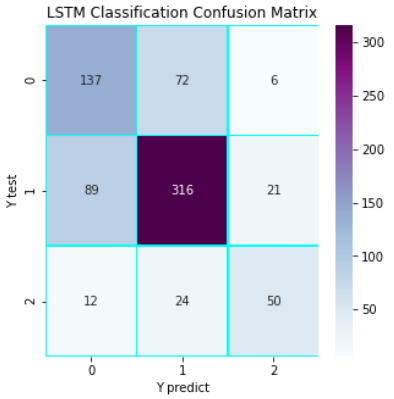

Figure 5. LSTM Classification Confusion Matrix.

Through the analysis of confusion matrix, this paper can observe the classification performance of the model in different emotional categories in detail. The confusion matrix shows that the model performs best in the category of neutral emotions, and accurately predicts 316 samples, which shows that the model has the strongest ability to identify neutral emotions. For positive emotions, the model correctly predicted 137 samples, but 72 samples were wrongly classified as neutral emotions and 6 samples were misjudged as negative emotions. In the classification of negative emotions, the model correctly identified 50 samples, but 24 samples were misjudged as neutral emotions and 12 samples were misjudged as positive emotions. These results show that the model is more challenging than the classification of neutral emotions when dealing with polar emotions (positive and negative).

 doi:10.20944/preprints202506.0968.v1

As shown in Table 1, the model performs best in neutral sentiment recognition, achieving an F1-score of 89.5%, which is consistent with the analysis from the confusion matrix. For positive sentiment, the model also achieves good performance with a precision of 83.5%. In comparison, negative sentiment recognition proves more challenging, with all metrics lower than the other two categories. The model achieves an average F1-score of 81.4% across all categories, demonstrating strong classification performance.

 doi:10.20944/preprints202506.0968.v1

Table 1. Performance Metrics for Different Sentiment Categories.

label	Precision(%)	Recall(%)	F1 (%)	Sample	label	Precision(%)	Recall(%)	F1 -score(%)
Positive	83.5	79.2	81.3	215	Positive	83.5	79.2	81.3
Negative	75.8	71.4	73.5	86	Negative	75.8	71.4	73.5
Neutral	88.7	90.3	89.5	426	Neutral	88.7	90.3	89.5

Overall, the experimental results show that the multi-modal federated learning model proposed in this paper has achieved satisfactory performance in the task of financial sentiment analysis. The model not only shows good convergence and generalization ability, but also reaches an acceptable accuracy level in all emotional categories.

V. Conclusions

This study aims to address the growing challenge of accurate sentiment analysis in the financial domain by developing a privacy-preserving machine learning model capable of handling multimodal data. Traditional centralized models face limitations in cross-institutional collaboration and data privacy protection. To overcome these barriers, the researchers propose a financial sentiment analysis framework that combines federated learning with multimodal data processing, specifically integrating text and voice inputs. The primary objective of this research is to enhance sentiment classification performance while maintaining data confidentiality across multiple institutions.

Through data analysis of 4,846 paired text-speech financial samples, the study identified three key findings: first, the model demonstrates strong overall sentiment classification accuracy, particularly for neutral sentiments; second, the dual-branch architecture effectively extracts complementary features from text and speech using BERT and Transformer-based encoders; third, the improved FedAvg algorithm ensures secure and efficient collaborative training across decentralized nodes. These findings suggest that combining multimodal learning with federated frameworks is a promising direction for financial sentiment analysis, offering both technical performance and compliance with data governance requirements.

The results of this study have significant implications for the field of financial artificial intelligence. Firstly, the proposed multimodal fusion method introduces a new perspective on how to jointly analyze text and speech in real-world financial sentiment tasks. Secondly, the use of federated learning challenges traditional assumptions about centralized model superiority, particularly in privacy-sensitive environments. Finally, the study opens new avenues for research into scalable, secure, and adaptive sentiment systems that can operate across institutions without compromising sensitive data.

Despite its contributions, the study has limitations. It exhibits reduced classification performance on polar emotions (positive and negative) compared to neutral sentiments, and the model's complexity introduces computational overhead. Future research could further explore advanced fusion mechanisms for better handling polar sentiment differentiation and optimize federated training algorithms to improve scalability and reduce training time across heterogeneous nodes. In conclusion, this study, through the integration of BERT-based multimodal feature extraction and federated learning, reveals a novel and effective framework for financial sentiment analysis. It provides new insights into building intelligent, privacy-aware systems for interpreting investor and analyst emotions, contributing both theoretical and practical advancements in the field of financial data science.

 doi:10.20944/preprints202506.0968.v1

References

1. Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782-796.
2. Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.
3. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.
4. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.
5. Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., ... & He, B. (2021). A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 35(4), 3347-3366.
6. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), 423-443.
7. Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. IEEE transactions on affective computing, 14(1), 108-132.
8. Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L. P., & Hoque, E. (2020). Integrating multimodal information in large pretrained transformers. In Proceedings of the conference. Association for computational linguistics. Meeting (Vol. 2020, p. 2359).