Natural Language Processing Journal 11 (2025) 100148

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Financial sentiment analysis for pre-trained language models incorporating dictionary knowledge and neutral features

Yongyong Sun, Haiping Yuan *, Fei Xu

Xi'an Technological University, School of Computer Science and Engineering, Xi'an, 710021, China

ARTICLE INFO

Keywords:
Financial sentiment analysis
BERT
Dictionary knowledge embedding
Ablation analysis
Neutral sentiment recognition

ABSTRACT

With increasing financial market complexity, accurate sentiment analysis of financial texts has become crucial. Traditional methods often misinterpret financial terminology and show high error rates in neutral sentiment recognition. This study aims to improve financial sentiment analysis accuracy through developing EnhancedFinSentiBERT, a model incorporating financial domain pre-training, dictionary knowledge embedding, and neutral feature extraction. Experiments on the FinancialPhraseBank, FiQA and Headline datasets demonstrate the model's superior performance compared to mainstream methods, particularly in neutral sentiment recognition. Ablation analysis reveals that dictionary knowledge embedding and neutral feature extraction contribute most significantly to model improvement.

1. Introduction

At the heart of financial markets lies the transmission and feedback of information. Each day, vast amounts of financial text are generated through news reports, analyst reports, company announcements, and social media discussions (Du et al., 2024). These texts not only reflect market dynamics but also influence investor sentiment and decision-making, significantly impacting asset prices and market trends. With the rise of the internet and social media, financial information has grown exponentially, exceeding manual analysis capabilities. Financial texts have distinct characteristics, including frequent jargon and implicit market signals, with subtle correlations between textual sentiment and market reactions. In response, Financial Sentiment Analysis has emerged, aiming to automatically extract sentiment information from massive financial texts to support investment decisions and market forecasts (Kearney and Liu, 2014).

This study aims to enhance the accuracy of financial text sentiment analysis, which holds both theoretical and practical importance. Theoretically, it promotes the application of natural language processing within the financial domain, addresses the unique challenges of semantic understanding in financial texts, and advances research on the expression and transmission mechanisms of sentiment information in this field. Practically, accurate sentiment analysis aids investors in interpreting market sentiment, supports systematic investment decisions, and helps regulators detect abnormal sentiment to maintain market stability. Therefore, improving the accuracy of financial text sentiment analysis not only advances research in financial NLP but also provides essential technical support for real-world financial applications (Agarwal, 2023).

Existing approaches to financial text sentiment analysis face significant technical challenges in practical applications. While pre-trained language models excel in capturing general linguistic representations, they lack a deep understanding of financial terminology and specific expressions (Mishev et al., 2020). For instance, terms like "quantitative easing" may suggest positive market expectations, while the sentiment of "debt restructuring" depends heavily on context—semantic nuances that general-purpose models struggle to interpret. Additionally, traditional models fall short in handling the implicit sentiment expressions typical of financial texts (Du et al., 2023). Unlike general texts that often use overt emotional language, financial texts employ professional, objective phrasing, with emotional cues subtly embedded, such as changes in the wording of a central bank's policy statement hinting at shifts in policy stance. Furthermore, existing methods struggle to effectively incorporate financial domain expertise, making it challenging to leverage external knowledge such as financial dictionaries or market reactions for sentiment analysis, thereby limiting the model's ability to fully comprehend complex financial texts (Ahmad and Umar, 2023).

This study proposes the EnhancedFinSentiBERT model to address challenges in financial text sentiment analysis through three key components: financial domain pre-training, dictionary knowledge integration, and neutral feature extraction. First, a large financial corpus is used to pre-train BERT with domain adaptation, enhancing its understanding of financial terminology and expressions. Second, the financial

* Corresponding author.
E-mail addresses: sunyongyong@xatu.edu.cn (Y. Sun), Yuanhaiping@xatu.edu.cn (H. Yuan).

https://doi.org/10.1016/j.nlp.2025.100148
Received 20 February 2025; Received in revised form 7 April 2025; Accepted 18 April 2025

domain lexicon ML, containing sentiment tendencies, frequencies, and weights of words, is employed to compute sentiment strength and scores, which are integrated into the model's architecture to better capture subtle sentiments in financial terms. Third, a neutral feature extractor is designed to identify and process the prevalent neutral expressions in financial texts, based on extensive analysis of neutral financial statements. This integrated approach enables the EnhancedFinSentiBERT model to accurately analyse complex financial sentiments while maintaining efficiency with limited labelled data, making it a robust and flexible tool for financial text analysis and decision-making.

The structure of this paper is as follows: Section 2 reviews related work and current limitations in financial sentiment analysis. Section 3 details the EnhancedFinSentiBERT model's architecture, including pre-training, dictionary integration, and neutral feature extraction. Section 4 provides pseudocode implementation. Section 5 describes the experimental setup, and Section 6 presents results and ablation studies. Finally, Section 7 summarizes findings and discusses limitations and future research directions.

2. Related work

This section reviews the established research on sentiment analysis in finance, explores the limitations of the current methodology, and describes the innovations of this study.

2.1. Financial sentiment analysis

Financial Sentiment Analysis, a critical branch of Sentiment Analysis, has gained significant attention from academia and industry in recent years. Its goal is to extract and quantify sentiment information from financial texts to support investment decisions and market forecasts. Research methods have evolved from traditional approaches to deep learning and pre-trained language models, with each stage yielding significant performance improvements (Cam et al., 2024; Man et al., 2019).

Early research relied on lexicon-based methods and traditional machine learning techniques. For instance, the financial sentiment lexicon by Loughran and McDonald (2011) improved the applicability of lexicon-based approaches by calculating sentiment tendencies based on the frequency of positive and negative words in financial texts (Loughran and McDonald, 2011). Feature engineering combined with classification algorithms, such as Support Vector Machines (SVMs), was also widely adopted. However, these methods struggled to capture complex linguistic structures and contextual nuances, especially in financial terminology (Kaur and Sharma, 2023).

The advent of deep learning revolutionized financial sentiment analysis. Ding et al. (2015) introduced deep learning techniques to financial news analysis by combining word embeddings with neural tensor networks, marking a turning point in the field (Ding et al., 2015). Subsequently, models like LSTM, CNN, and doc2vec have been applied, with CNN models significantly outperforming both traditional and other deep learning methods in specific tasks, such as StockTwits sentiment analysis (Sohangir et al. 2018) (Sohangir et al., 2018). Despite their success in capturing semantics and long-distance dependencies, deep learning models often require large amounts of labelled data, which remains a challenge in the resource-constrained financial domain.

The emergence of pre-trained language models, such as BERT, has further transformed financial sentiment analysis (Devlin et al., 2019). These models, pre-trained on large-scale corpora and fine-tuned for specific tasks, offer a richer understanding of context and semantics while reducing the need for extensive labelled datasets (Mishev et al., 2020). Domain-specific adaptations, such as FinBERT, enhance BERT by pre-training on financial corpora, significantly improving financial text analysis accuracy. Such models address the limitations of labelled data availability in the financial field.

To enhance performance further, researchers have incorporated domain knowledge into models. For example, Yang et al. (2020) proposed FinBERT-Tone, which integrates financial lexicon information to improve sentiment analysis (Yang et al., 2020). Choe et al. (2023) introduced FiLM, a financial pre-trained model using a diverse corpus, achieving superior performance on various financial tasks while reducing energy consumption by 82% (Choe et al., 2023). These studies underscore the importance of corpus diversity and domain knowledge fusion, demonstrating the potential of combining data-driven approaches with financial expertise to enhance model performance and interpretability.

2.2. Limitations of existing methods and innovations in this study

Despite significant progress in Financial Sentiment Analysis (FSA), existing approaches face unresolved challenges. Pre-trained language models, while effective at capturing general linguistic representations, lack domain-specific knowledge, limiting their understanding of financial terminology and expressions. Additionally, attempts to integrate domain-specific knowledge during fine-tuning remain insufficient, particularly in leveraging lexical knowledge within the financial domain (Karanikola et al., 2023). Traditional sentiment analysis methods also focus primarily on positive and negative sentiments, neglecting the effective handling of neutral expressions, which are both prevalent and challenging to define in financial texts (Lin and Liao, 2024).

To address these issues, this study proposes several methods. First, the BERT model is pre-trained on a financial corpus, enhancing its comprehension of financial texts. Second, financial domain dictionary knowledge is incorporated, improving the model's ability to capture subtle financial sentiment and enhancing its interpretability. Finally, a neutral feature extractor is introduced to handle neutral expressions common in financial texts, enabling the model to better distinguish between subtle sentiment differences, such as slightly positive, neutral, and slightly negative.

The method in this study has achieved significant performance improvements on the PhraseBank, FiQA Task 1, and Headline benchmark datasets, demonstrating its effectiveness and applicability in the field of financial text analysis. These improvements provide new ideas and tools for sentiment analysis in the financial sector, with the potential to play an important role in practical applications such as investment decision support and market sentiment analysis.

3. Construction of the EnhancedFinSentiBERT model

This section details the overall architecture of the proposed model and its various components. The approach in this study is based on the BERT model, which is used to improve the performance of the model on financial text analysis tasks through financial domain pre-training, financial lexicon knowledge incorporation, and neutral feature extraction techniques.

3.1. Overview of the model

The EnhancedFinSentiBERT model integrates financial domain pre-training, lexical knowledge incorporation and neutral feature extraction components in order to improve the accuracy of financial text sentiment analysis. The overall architecture of the model is shown in Fig. 1 below. This study proposes a three-branch fusion architecture for financial sentiment analysis, consisting of a financial pre-training branch, a lexical feature enhancement branch, and a neutral feature processing branch. The financial pre-training branch employs NLTK for sentence disambiguation, a 15% random mask MLM strategy for pre-training, and a BERT feature extractor to obtain deep semantic representations. The lexical feature enhancement branch refines word features through an initial embedding layer, an MLP feature mapping layer, and a multi-head attention mechanism. The neutral feature processing branch

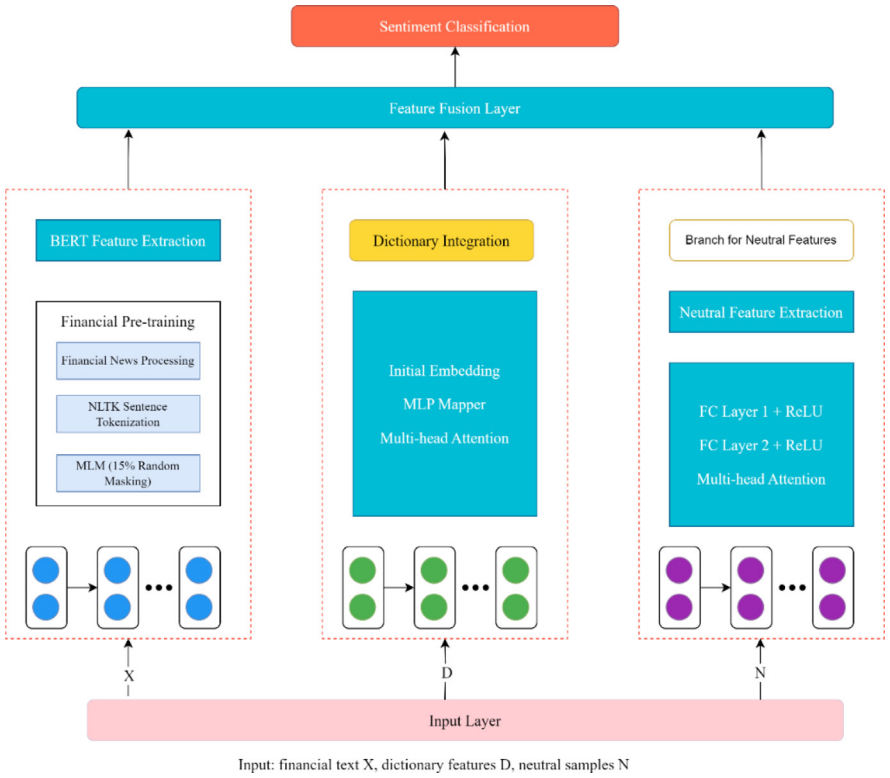

Fig. 1. Overall framework diagram.

uses a feature extractor, a two-layer ReLU-activated fully connected network, and a multi-head attention mechanism to effectively capture neutral semantic information in the text.

These three functional branches first receive the corresponding feature inputs at the input layer, and form a feature sequence representation after their respective feature extraction and transformation. Specifically, the blue sequence $\{x_1, x_2, \ldots, x_n\}$ represents the deep semantic features of the text extracted by BERT, the green sequence $\{d_1, d_2, \ldots, d_m\}$ shows the domain feature representation of the lexical knowledge after embedding and mapping, and the purple $\{n_1, n_2, \ldots, n_k\}$ sequence reflects the transformation state of neutral features in the processing branch. These feature sequences achieve effective integration of multi-source information through the feature fusion layer, and finally the sentiment prediction results are output by the classification layer. The architecture organically combines text-based semantic features, domain-specific knowledge and neutral expression features to construct a multi-dimensional feature representation framework.

Each of the key components of the model, including the BERT model, the Financial Domain Corpus, the Financial Lexicon Knowledge Incorporation and the Neutral Feature Extractor, are described in detail below.

3.2. Theoretical foundation of component integration

The EnhancedFinSentiBERT model uses three complementary components addressing key financial sentiment analysis challenges. Domain adaptation helps with specialized terminology, knowledge enhancement improves emotional nuance detection, and boundary optimization better identifies neutral statements. This multi-dimensional approach overcomes limitations of single-method solutions for analysing complex financial texts.

Financial domain pre-training achieves domain adaptation of the model, which from an information theory perspective can be viewed as reducing the conditional entropy of the model for financial language, enabling it to build more accurate financial text representations. Financial texts have unique linguistic features, such as professional terminology (“quantitative easing”, “debt restructuring”) and specific expressions, which general language models find difficult to accurately understand. Through pre-training on large-scale financial corpora, the model adjusts its internal representation to adapt to financial contexts, reducing the semantic gap between general language understanding and financial professional expressions, laying the foundation for subsequent tasks.

The dictionary knowledge embedding component adopts a hybrid intelligence approach, combining the implicit neural representations of pre-trained models with explicit symbolic knowledge provided by dictionaries. This fusion can be represented as an adaptive weighting process, where weights are dynamically learned through a multi-head attention mechanism. The sentiment intensity and word frequency information provided by financial dictionaries become the most influential features, playing an important guiding role in the model’s sentiment judgement. Purely data-driven approaches often lack an understanding of market reactions when dealing with financial terminology, while ML dictionaries learn word sentiment tendencies directly through market reactions, providing explicit sentiment knowledge that is difficult to acquire through data-driven methods. This external knowledge complements the implicit patterns learned by pre-trained models, providing richer judgement bases for sentiment analysis, while improving the model’s interpretability.

The neutral feature extractor optimizes the model’s ability to recognize neutral expressions through a specialized neural network structure. This component transforms BERT’s output features into representations more suitable for identifying neutral sentiment through feature transformation and processing. Neutral expressions in financial texts often contain subtle market signals, with blurred boundaries from weak sentiment expressions. Through a specialized neutral feature learning mechanism, the model can more precisely capture the characteristics of neutral sentences, reducing confusion with weak sentiment expressions. This mechanism effectively reduces the misclassification of neutral sentiment, especially addressing the confusion between neutral and positive sentiment.

Table 1
Distribution of corpus data in the financial domain.

Group	Classification	Descriptions	Data size
News	Currency	Includes coverage of exchange rate movements, central bank policies and the impact of international trade on currencies	60.25M
	Commodity	Price and supply/demand analyses covering commodities such as crude oil, precious metals and agricultural products	190.7M
	Economics	Coverage of macroeconomic indicators and policy trends such as GDP, employment rate, inflation, etc.	196.1M
	Economic indicators	Tracking various economic indices such as PMI, Consumer Confidence Index, etc.	68.4M
	Stock Market	Coverage of major global stock indices, individual stock performance and market trend analysis	340.1M
	cryptocurrency	Coverage of price volatility, regulatory policy and technological developments in cryptocurrencies such as Bitcoin	100.6M
Analysts Analysis	personal analysis	Financial analysts' personal insights and forecasts on market trends, investment opportunities and risks	70.6M

These three components form complementary reinforcing relationships at the theoretical level, constituting a collaborative system. The domain adaptation mechanism adapts the model to financial language characteristics, the knowledge enhancement mechanism introduces professional financial sentiment knowledge, and the neutral feature extraction mechanism improves the recognition precision of neutral expressions. From the perspective of ensemble learning theory, these three mechanisms target different types of errors, producing complementary error correction effects: domain pre-training enhances financial context adaptability but may not precisely capture the sentiment tendencies of terminology, dictionary knowledge compensates for this deficiency; dictionary knowledge provides terminology sentiment tendencies but lacks context sensitivity, while the pre-trained model possesses this capability; and the neutral feature extractor specifically optimizes the ability to distinguish between neutral and weak sentiment expressions.

3.3. BERT

BERT is a deep bidirectional language representation model based on Transformer, pre-trained on large-scale unlabelled text corpus using two unsupervised tasks: masked language modelling (MLM) and next-sentence prediction (NSP). Its input representation combines token, positional, and segmental embeddings, with special [CLS] and [SEP] tokens for handling various NLP tasks. For sentence-pair tasks, segment embeddings distinguish between first and second sentences, while the [CLS] token's final hidden state represents the entire sequence. This study employs BERT-base (110M parameters) instead of BERT-large (340M parameters) to balance performance and computational efficiency, as BERT-base has demonstrated strong performance across NLP tasks while maintaining faster processing speeds, which is crucial for analysing large volumes of financial texts.

3.4. Financial domain corpus

This study uses a self-constructed financial domain corpus to adapt BERT to financial language characteristics. The corpus includes articles from financial news outlets and analyst reports from 2010 to September 2024, comprising approximately 1 GB of text and over 9 million sentences. Table 1 shows the corpus data distribution:

During pre-training, this study adopts the same Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks as the original BERT, with adjustments to training parameters and strategies to accommodate the larger dataset. The batch size was set to 64, and gradient accumulation was used to effectively expand the training batch size, enhancing model stability. The learning rate was initialized at 2e-5 and followed a linear decay strategy. To improve training efficiency, fp16 precision training was applied, and an initial pre-training period of 30,000 steps was set. The entire pre-training process lasted approximately 142,000 steps to fully leverage the large-scale dataset.

The training process of the model is shown through the loss profile graph in Fig. 2 and the loss decline rate graph in Fig. 3, with data recorded every 300 steps. The initial loss begins at 2.8025 and decreases rapidly during the first 20,000 steps, showing large fluctuations in the decline rate and a high average value, indicating a rapid learning phase. After this, the loss continues to decline smoothly, with reduced fluctuations in the decline rate, which remains positive, signifying effective ongoing learning. In the later stages, after 80,000 steps, the decline rate approaches zero, reflecting a significant slowdown in learning. By the end of training, around 140,000 steps, the loss reduces to approximately 1.3510, representing an overall reduction of over 50 percent. Small fluctuations are present throughout, but there are no signs of overfitting, and the negative values in the decline rate graph indicate normal random variations during training.

The pre-training results show that the strategy of this study is effective. The model went through the typical process of fast learning, steady learning and finally convergence, indicating that it adapted to financial text features. The slow decrease in loss at the late stage of training indicates that the model is still learning subtle features, which is crucial for processing complex financial texts. The absence of significant overfitting indicates that the dataset is appropriately sized for the model's generalization ability.

Pre-training in the financial domain significantly improves the model's understanding of financial texts by exposing it to a large number of diverse financial texts. This process enabled the model to not only learn jargon and industry-specific expressions, but also adapt to contexts and expressions specific to the financial domain. The model potentially acquires knowledge of basic financial concepts and market dynamics while processing a large-scale financial corpus, and at the same time improves its comprehension of the time series of financial events. Through exposure to different types of financial texts, such as news and analytical reports, the model learns to understand multiple forms of expression for the same concept and is better able to capture obscure expressions and subtle messages commonly found in financial texts. This comprehensive learning process enabled the model to more accurately understand text content and sentiment tendencies in subsequent financial sentiment analysis tasks, laying the foundation for further task-specific fine-tuning.

3.5. Integration of financial lexicon knowledge

This research adopts the ML lexicon proposed by García et al. (2023) (Garcia et al., 2023). Unlike the manual management method used by Loughran and McDonald (2011), which relies on static word lists defined by experts, the ML lexicon employs a data-driven approach. By applying the robust multinomial inverse regression (robust MNIR) model to 85,530 earnings call transcripts, Garcia et al. associated the occurrence of n-grams with contemporaneous stock price reactions, thus establishing the emotional polarity of words. This

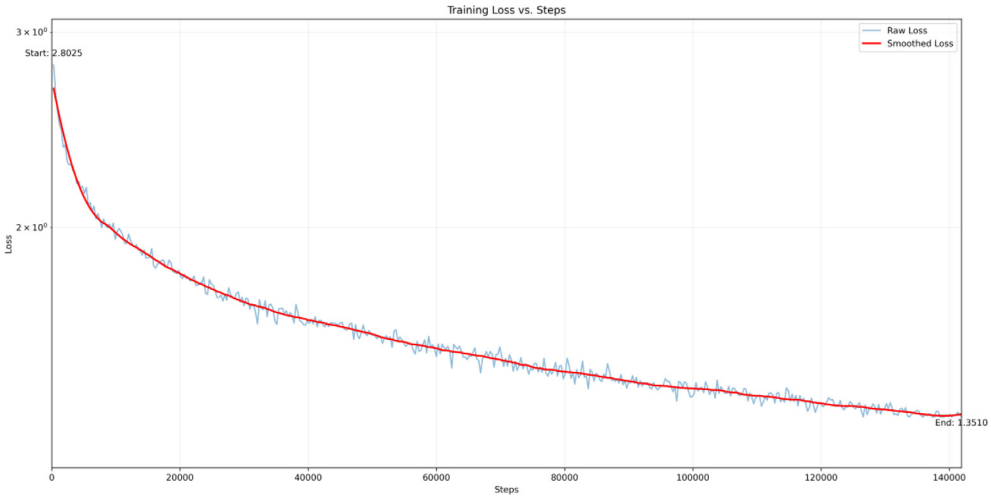

Fig. 2. Training loss graph.

Fig. 3. Plot of decline rate of training losses.

market-supervised method can capture contextual nuances often overlooked by human coders. For example, while the LM lexicon explicitly labels "problem" as negative, the ML algorithm indicates that its actual emotional impact depends on contextual combinations such as "technical problem" (negative) versus "stock issuance" (neutral), as demonstrated by their differential predictive ability in stock return regressions.

To fully utilize this resource, the current research processed and enhanced the lexicon in multiple ways. First, word frequency information provided in the original lexicon was transformed through a logarithmic scale to calculate word weights and was normalized to balance the influence of high and low-frequency words. The positive, negative, and neutral sentiment probabilities from the original lexicon were retained and used to calculate additional metrics, including sentiment polarity and sentiment intensity. These metrics provide insights into the directional tendency and emotional strength of each word. Finally, a comprehensive sentiment score was calculated that fuses sentiment polarity, intensity, and normalized word weights. This score simultaneously reflects both the sentiment tendency and importance of words in the corpus, providing a comprehensive metric for sentiment analysis.

Compared to existing financial lexicon integration methods, this research's implementation adopts several different approaches: First, it introduces a dynamic weight adjustment mechanism that adjusts influence weights based on word performance in different financial contexts, rather than using static weights; second, it adopts multidimensional sentiment representation, considering not only the positive and negative polarity of words but also quantifying their intensity and degree of market impact; finally, it implements a context-sensitive fusion strategy, enabling lexicon knowledge to dynamically interact with BERT's contextual representations through a multi-head attention mechanism, thereby enhancing the model's ability to capture subtle sentiment expressions in financial texts.

The importance of each lexicon feature in the model is shown in Fig. 4. Among these features, sentiment intensity scores highest at 0.8442, highlighting the critical role of sentiment intensity in financial contexts. Word frequency and its derivative indicators, including weight and normalized weight, follow closely with a score of 0.8091, confirming the importance of word usage frequency. The comprehensive sentiment score is 0.7213, demonstrating the effectiveness of integrating sentiment factors and word importance. Sentiment polarity, though scoring lower, still makes a meaningful contribution to the model.

The distribution of feature importance highlighted above emphasizes the advantages of the lexicon model in measuring word importance, especially when each feature is considered in combination, making the lexicon more effective. In order to effectively utilize these rich lexical features in the model, this study designs a lexical embedding module which is integrated in parallel with the pre-trained language model. Fig. 5 illustrates the architecture of the financial sentiment BERT model incorporating domain knowledge. In this study,

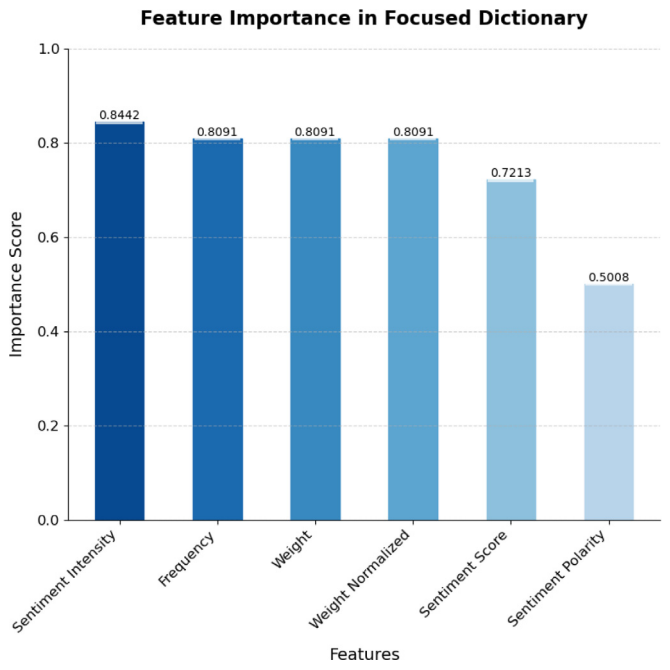

Fig. 4. Importance map of dictionary features.

a pre-trained BERT model in the financial domain is used to generate deep text representations. For the financial sentiment analysis task, the input is in the format '[CLS] Financial Text [SEP]'. The BERT model outputs a representation of the entire sequence as shown in Eq. (1):

$$H = BERT\,(x1, x2, \ldots, xn) \tag{1}$$

where $H \in \mathrm{R}^{n\times d}$, n is the sequence length and d is the hidden dimension of BERT.

For each word xi in the input text, this study looks up its corresponding feature vector $D(xi)$ from the processed dictionary. If a word is not in the dictionary, a zero vector is returned. The lexicon embedding module first creates the initial embedding E, and then maps it to a BERT-compatible embedding space by means of a multilayer perceptron (MLP), as shown in Eqs. (2)–(3):

$$E_{initial} = Embedding\left(D\left(x_1\right), D\left(x_2\right), \ldots, D\left(x_n\right)\right) \tag{2}$$

$$E_{compatible} = MLP\left(E_{initial}\right) \tag{3}$$

where the mathematical definition of MLP is shown in Eq. (4):

$$MLP(x) = W_k(\sigma(\ldots W_2(\sigma(W_1 x + b_1)) + b_2)\ldots) + b_k \tag{4}$$

σ is the activation function, which is the ReLU function chosen for this study, and Wi and b_i are the weights and bias of layer i.

At the core of the lexicon embedding module is a multi-head attention structure. This mechanism is used in this study to dynamically adjust the importance of different words and features to enable the interaction of lexical knowledge with the BERT contextual representation. The multi-head attention mechanism is defined as shown in Eqs. (5)–(7):

$$MultiHead\,(Q, K, V) = Concat\left(head_1, \ldots, head_h\right) W^O \tag{5}$$

$$where\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{6}$$

$$Attention\,(Q, K, V) = softmax\left(\frac{(QK^T)}{\sqrt{d_k}}\right)V \tag{7}$$

where Q, K, V are the query, key, and value matrices, respectively, d_k is the dimension of attention, and h is the number of attention heads.

The output H of BERT is used as contextual information in the model to interact with $E_{compatible}$ to obtain a context-aware lexical representation. The multi-head attention implementation provided by PyTorch is used, which handles the query, the conversion of keys and values, and the computation of attention at the bottom layer as shown in Eq. (8):

$$E_{contextual} = MultiHead\left(H, E_{compatible}, E_{compatible}\right) \tag{8}$$

where $E_{contextual}$ denotes the context-aware dictionary representation after multi-head attention processing.

Finally, the output H of BERT is fused with the contextual representation of the dictionary and computed as shown in Eq. (9):

$$F = Fusion\left(H, E_{contextual}\right) \tag{9}$$

$Fusion$ employs a linear layer to combine the output of BERT with the context-aware dictionary representation. It first splices H and $E_{contextual}$ in feature dimensions and then obtains the fusion representation by linear transformation as shown in Eq. (10):

$$Fusion = Linear\left(Concat\left(H, E_{contextual}\right)\right) \tag{10}$$

where $Concat$ denotes the splicing operation on the feature dimension.

3.6. Neutral feature extractor

Financial texts contain a large number of neutral expressions that may contain market signals, and the identification and processing of neutral expressions is an important part of financial sentiment analysis. The error analysis of the preliminary experiments shows that the prediction bias of the model mainly appears in the judgement of neutral sentiment. The normalized confusion matrix in Fig. 6 visualizes this phenomenon.

As can be seen from the figure, the main misjudgements of the model are concentrated between neutral emotions and other emotion categories. Of the 148 positive samples, 27.7% were misclassified as neutral; of the 561 neutral samples, 7.3% were misclassified as positive; and of the 261 negative samples, 3.8% were misclassified as neutral. These three types of errors totalled 92 samples, representing 9.5% of the total sample size of 970, but 79.3% of the 116 total misclassifications. This phenomenon highlights the limitations of the model in distinguishing between neutral and non-neutral emotions, especially when dealing with subtle emotional expressions. For example, 'sponda will record a profit from the sale of 8.5 mln euro ($12.4 mln).' A positive statement like this was misjudged as neutral, while a neutral statement like 'as of august 2008, glaston's north asian sales and service region is upgraded to a new market area, north asia.' was misjudged as positive. This pattern of misclassification not only reflects the complexity of sentiment expression in financial texts, but also highlights the importance of improving the accuracy of the model in recognizing neutral sentiment.

To address this problem, this study designed neutral feature extractors that enhance the understanding of neutral expressions.

3.6.1. Structure of the neutral feature extractor

In order to break through the existing limitations of neutral emotion recognition, a neutral feature extractor is designed in this study. The neutral feature extractor adopts a lightweight but effective neural network structure, which mainly consists of an input layer, two fully-connected layers, ReLU activation function, multi-head attention layer and average pooling layer. The structure is shown in Fig. 7. The structure begins by receiving the 768-dimensional hidden state representation from the BERT model, which is reduced to 384 dimensions through a fully connected layer followed by a ReLU activation function for nonlinear transformation. A second fully connected layer then refines features while maintaining 384 dimensions. Next, a multi-head attention layer with 4 heads captures neutral feature information at various scales. Finally, the attention outputs are aggregated using

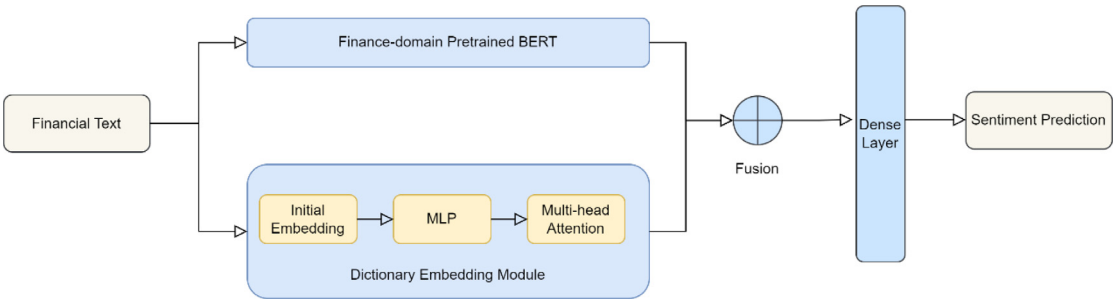

Fig. 5. Architecture of financial sentiment BERT model incorporating domain knowledge.

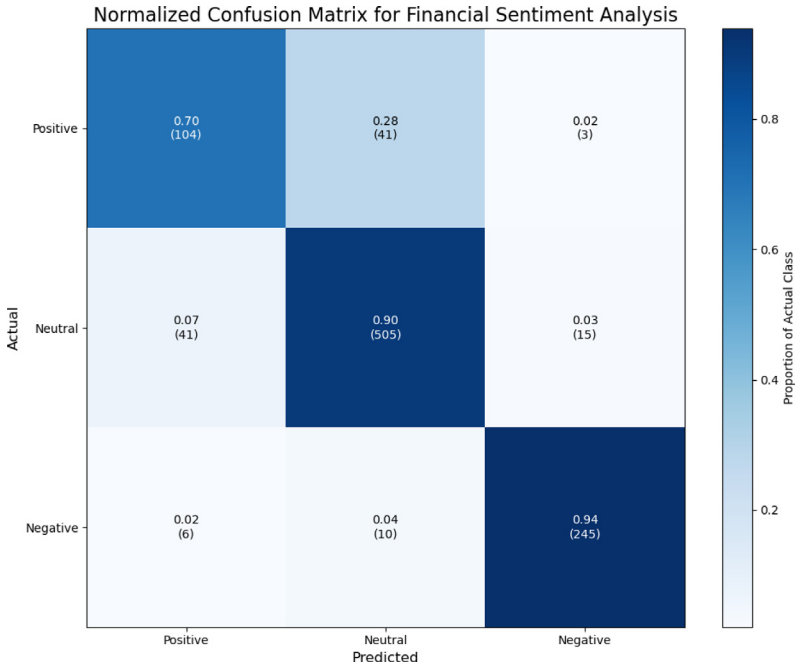

Fig. 6. Normalized confusion matrix.

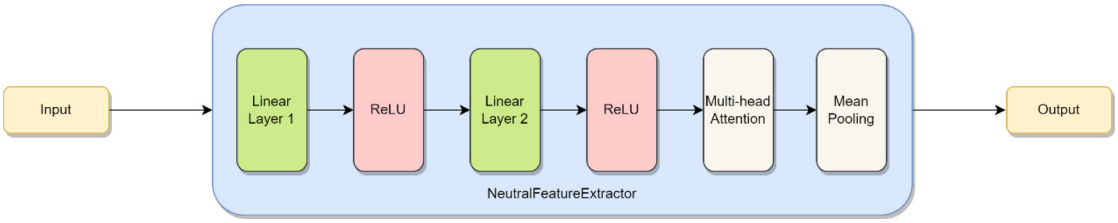

Fig. 7. Neutral feature extractor.

average pooling to produce the final neutral feature representation. This process is mathematically described in Eqs. (11)–(14).

$$h_1 = ReLU\left(W_1 \cdot BERT\left(x\right) + b_1\right) \tag{11}$$

$$h_2 = ReLU\left(W_2 \cdot h_1 + b_2\right) \tag{12}$$

$$h_{att} = MultiHead\left(h_2, h_2, h_2\right) \tag{13}$$

$$f_{neutral} = MeanPooling\left(h_{att}\right) \tag{14}$$

where x is the input text, W_1, W_2, b_1, b_2 are the learnable parameters, and $f_{neutral}$ is the final neutral feature representation.

The neutral feature extractor harnesses BERT's contextual representations, employing feature downscaling and extraction through fully connected layers, while utilizing multi-attention mechanisms to capture neutral features at various scales. These features are consolidated into a unified vector representing neutral characteristics of the input text through average pooling operations. This architectural design enhances the model's capability to identify subtle neutral expressions and differentiate them from slightly positive or negative sentiments within financial texts. Through this approach, the feature extractor strengthens the model's perception of nuanced semantic differences, establishing a robust foundation for subsequent sentiment analysis tasks.

3.6.2. *Dataset used to extract neutral features*

In order to train and optimize the neutral feature extractor, this study utilizes the SEntFiN 1.0 (Sentiment Analysis for Financial News) dataset. The SEntFiN 1.0 dataset contains 10,753 financial news headlines, each of which is annotated with the entities mentioned therein and their corresponding sentiment tendencies (Sinha et al., 2022). The

dataset includes a sample of 5,074 positive sentiments, 5,517 neutral sentiments, and 3,814 negative sentiments. Below is an example of the dataset:

"S No.": 8,
"Title":"HOEC could retest 30-35 levels: Ashwani Gujral",
"Decisions": ['HOEC': 'neutral'],
"Words":7

In this study, special attention is given to the 5,517 neutral samples in the SEntFiN dataset. These samples, consisting of typical objective statements, factual reports, and neutral expressions in financial texts, provide valuable learning materials for the neutral feature extractor. However, this dataset alone is insufficient for the model to fully learn neutral features. To address this, a large number of neutral-objective articles are selected from professional financial news websites such as Reuters, CNBC, Investing.com, and Seeking Alpha. Neutral sentences are extracted from these articles to ensure they are free from emotional bias.

In practice, the SEntFiN dataset is first preprocessed to extract all the samples labelled as neutral, and then merged with the extracted neutral sentiment data, and these samples are used to train the neutral feature extractor, with the goal of enabling the extractor to accurately identify and represent neutral features in financial texts, especially in the presence of multiple entities and potentially conflicting sentiments.

4. Algorithm pseudo-code

Algorithm 1 presents the complete training process of EnhancedFinSentiBERT, which integrates pre-trained language models with domain-specific knowledge for financial sentiment analysis. The algorithm requires three key inputs: financial text corpus D containing labelled training data, financial dictionary Dict with domain-specific terminology, and neutral text samples N for distinguishing neutral sentiments. Combined with standard parameters like learning rate η and maximum iterations T, these inputs enable the model to effectively analyse sentiment in financial texts across all sentiment categories.

Algorithm 1: EnhancedFinSentiBERT Training Process

Input: Financial text corpus D, Financial dictionary $Dict$, Neutral text samples N, Learning rate η, Maximum iterations T

Output : Trained model parameters W_T

Initialize BERT pre-trained weights to obtain W_0
Fine-tune using financial corpus D to get $W_0' = FinBERT(W_0, D)$
Initialize dictionary embedding matrix $E \in \mathbb{R}^{d \times |V|}$ using $Dict$
Initialize neutral feature extractor M using samples N
while t < T **do**
  for $(x_i, y_i) \in D$ do
    $H = Tokenize(x_i)$ ▷ Text tokenization
    $F = FinBERT(H)$ ▷ BERT features
    $E_i = DictEmbed(H, Dict)$ ▷ Dictionary features
    $R = Fusion(F, E_i)$ ▷ Feature fusion
    $N = NeutralExtract(H)$ ▷ Neutral features
    $Z = Fusion(F, E_i, N)$ ▷ Feature concatenation
    $y_{pred} = softmax(W_i \cdot Z + b)$ ▷ Sentiment prediction
    $L = CrossEntropy(y_{pred}, y_i)$ ▷ Compute loss
    $W_i + 1 = AdamW(W_i, \nabla L, \eta)$ ▷ Parameter update
  end for
end while
return W_T

5. Experimental setup

5.1. Dataset

This research evaluated the model on three widely used financial sentiment analysis datasets: Financial PhraseBank, FiQA Task 1, and Headline.

Financial PhraseBank contains 4,845 financial news headlines and phrases labelled by experts as positive, negative, or neutral, with an average length of 12.8 words (Malo et al., 2014). This dataset presents a natural imbalanced distribution, with neutral samples dominating, which reflects the characteristics of typical financial news. The evaluation used both the complete dataset and a high-consistency subset (2,264 samples with expert consensus).

The FiQA Task 1 dataset includes 1,174 finance-related social media posts and news headlines, with continuous sentiment scores ranging from −1 to 1 (Maia et al., 2018). These texts have an average length of 18.3 words, longer and more colloquial than Financial PhraseBank, and have a balanced distribution of positive, neutral, and negative sentiment.

The Headline dataset (Sinha and Khandait, 2020) consists of news headlines from the financial domain, with a particular focus on the gold commodity market. In this research, the dataset was used for sentiment analysis evaluation, which includes multiple classification subtasks, each corresponding to a judgement question about the gold market information (such as price movements, asset comparisons, etc.) (Sinha and Khandait, 2021).

5.2. Evaluation metrics

This paper uses three standard datasets to evaluate the performance of different model categories on financial sentiment analysis tasks. For the Financial PhraseBank (FPB) dataset, two versions were used: one with 50% expert annotation agreement and another with 100% complete expert consensus. For FiQA Task 1, following BloombergGPT (Wu et al. 2023), this paper converted continuous sentiment values into a classification task for evaluation (Wu et al., 2023). On the Headline dataset, the model's ability to classify sentiment in financial news headlines was evaluated. Weighted F1 score was used as the evaluation metric across all datasets.

5.3. Baseline models

In the experiments, this paper considered three categories of baseline models: general pre-trained models BERT-base and XLNeT; general large language models GPT-4 and Llama 2; and finance domain-specific models FinBERT and BloombergGPT. This classification covers basic pre-trained models, large-scale general language models, and models specifically optimized for the financial domain, providing a comprehensive comparison benchmark for this paper.

5.3.1. BERT-base

As a variant of the original BERT model, it has proven its strong performance and wide applicability in natural language processing tasks. The approach in this paper is based on the improvement of BERT-base-uncased, which enables a direct comparison of the performance gains brought about by the proposed improvements, thus better demonstrating the advantages of the improved model for financial sentiment analysis tasks.

5.3.2. XLNet

An important variant model of BERT, XLNet, was chosen as the baseline for this study. XLNet uses alignment language modelling to overcome some of the limitations of BERT, and performs well in dealing with long text and long distance dependencies. The selection of the variant model as a baseline helps to comprehensively evaluate the performance of the improved model in financial text sentiment analysis, as well as to explore the potential of different pre-training strategies for application in the financial domain.

Table 2
Performance of different model categories on financial sentiment analysis tasks.

Model name	Model category	FPB-50% agreement	FPB-100% agreement	FiQA task 1	Headline
BERT-base	Base PLMs	0.817	0.968	0.820	0.895
XLNeT		0.857	0.966	0.865	0.913
GPT-4	General LMs	0.830	0.960	0.872	0.860
Llama 2		0.860	0.970	0.820	0.942
FinBERT(Yang et al.)	Fin.-Specific	0.865	0.960	0.845	0.908
BloombergGPT		0.51	–	75.07	82.20
EnhancedFinSentiBERT		**0.870**	**0.980**	**0.880**	**0.976**

5.3.3. GPT-4

As a large language model developed by OpenAI, GPT-4 demonstrates excellent performance in general understanding and generation tasks through large-scale pre-training and Reinforcement Learning from Human Feedback (RLHF) techniques (Li et al., 2023). GPT-4 was selected as a baseline model to evaluate the performance of state-of-the-art general large language models on financial sentiment analysis tasks, and to explore their transfer learning capabilities without requiring specific optimization for the financial domain. Through comparison with domain-specific models, the performance of this paper can be more comprehensively evaluated.

5.3.4. Llama 2

Llama 2 is an open-source large language model developed by Meta. This model adopts innovative pre-training methods and optimized architecture design, showing excellent performance on general understanding tasks (Wang et al., 2023). Including Llama 2 in the baseline models for this paper helps to comprehensively compare performance differences of various types of large language models on financial sentiment analysis tasks, and to explore the capability comparison between open-source models and proprietary models.

5.3.5. Finbert

The FinBERT model (Yang et al.) is a finance domain-optimized model that has been pre-trained on various financial texts including 10-K reports, news, and tweets. These benchmarks can evaluate the effectiveness of EnhancedFinSentiBERT in financial domain adaptation and provide baselines for comparing different pre-training strategies.

5.3.6. BloombergGPT

BloombergGPT is a large language model developed by Bloomberg specifically for the financial domain. Through extensive training on financial news, reports, and market data, this model possesses powerful financial text understanding capabilities. Selecting BloombergGPT as a baseline model allows for evaluation of the comparative effectiveness between large language models specifically built for the financial domain and the method proposed in this paper on sentiment analysis tasks, thereby comprehensively validating the advantages and innovations of EnhancedFinSentiBERT in financial text sentiment analysis.

5.4. Experiment details

This experiment was conducted on an NVIDIA RTX 4090 GPU, using PyTorch 2.0.0 and Transformers 4.44.2 libraries. Key hyperparameters include: batch size 16, learning rate 2e-5, training epochs 5, maximum sequence length 128, and AdamW optimizer. These settings strike a balance between computational efficiency and model performance, providing an effective framework for fine-tuning pre-trained language models.

Regarding dataset partitioning, the FiQA Task 1 dataset was divided using a 10-fold partitioning method with 90% used for training and 10% for testing. Financial PhraseBank utilized both the complete dataset and the subset with complete expert consensus for evaluation. To avoid randomness, this study employed 10-fold cross-validation to evaluate model performance and reported average results. The Headline dataset was evaluated according to its original partition.

6. Experimental results and analysis

6.1. Overall performance comparison

Table 2 lists the results of EnhancedFinSentiBERT and benchmark methods on the three datasets: Financial PhraseBank, FiQA Task 1, and Headline.

Financial PhraseBank (FPB) As shown in Table 2, the EnhancedFinSentiBERT model performs optimally on the Financial PhraseBank dataset. On the complete dataset with 50% annotation agreement, the model achieved an F1 score of 87.0%, higher than all baseline models. On the 100% expert consensus subset, the model performed even better, reaching an F1 score of 98.0%, leading other models. Notably, compared to XLNeT's 85.7% and GPT-4's 83.0%, this model shows better results in processing specialized financial texts.

Experimental data shows that all models perform better on the expert consensus subset than on the complete dataset. EnhancedFinSentiBERT's F1 score on the consensus subset is 11 percentage points higher than on the complete dataset, increasing from 87.0% to 98.0%. Generally, all models show an F1 score improvement of 8.0 to 11.0 percentage points on the consensus subset compared to the complete dataset, indicating that samples with disagreement among financial experts indeed present greater classification challenges.

Notably, BloombergGPT, as a large language model specialized in finance, only achieved an F1 score of 51% on the FPB-50% dataset, significantly lower than other models. This suggests that even large models specifically trained for the financial domain may face challenges in sentiment analysis tasks, and that model architecture and task-specific optimization may be more important than relying solely on large-scale pre-training.

FiQA Task 1 Table 2 also lists the performance of various models on the FiQA Task 1 dataset. On this social media financial text dataset, EnhancedFinSentiBERT leads all baseline models with an F1 score of 88.0%. This result indicates that the model maintains stable performance when processing more informal financial texts containing numerous investor opinions.

Headline dataset: In the financial news headline classification task, the proposed model also stands out with an F1 score of 97.6%, surpassing all other baseline models. Llama 2 achieved an F1 score of 94.2% ranking second, XLNeT reached 91.3%, FinBERT 90.8%, BERT-base 89.5%, and GPT-4 86.0%, while BloombergGPT performed relatively weakly on this task at only 82.2%. These varied results indicate that models with different architectures have different strengths when processing brief, highly condensed financial texts.

Overall, EnhancedFinSentiBERT maintained good performance across all three datasets, demonstrating its good generalization capability. These results have important implications for practical applications in the financial domain, especially in analysing investor sentiment on social media and market prediction.

6.2. Analysis of ablation experiments

In order to fully evaluate the contribution of EnhancedFinSentiBERT model components, this study conducted a series of ablation experiments. The integration of dictionary knowledge and the neutral feature

Table 3
Performance of different model categories on financial sentiment analysis tasks.

Model name	FPB-50% agreement	FPB-100% agreement	FiQA task 1	Headline
BERT-base	0.817	0.968	0.820	0.895
BERT-Dictionary	0.856	0.970	0.842	0.925
BERT-Neutral	**0.870**	0.972	0.857	0.932
BERT-financial	0.837	0.967	0.832	0.912
BERT-financial-Dictionary	0.843	0.967	0.850	0.931
BERT-financial-Neutral	0.861	0.973	0.863	0.947
EnhancedFinSentiBERT	**0.870**	**0.980**	**0.880**	**0.976**

extractor are the core of the experiments, therefore separate functional validation of these components is required when conducting ablation experiments. The validation results from experiments conducted on the Financial PhraseBank, FiQA Task 1, and Headline datasets are shown in Table 3.

From Table 3, it can be observed that the baseline BERT-base model performs differently across the four evaluation sets, with its performance serving as an experimental control benchmark. By incorporating different functional modules, the performance of various variant models shows significant differences, reflecting the differentiated contributions of each component to the overall model performance.

Dictionary knowledge integration (BERT-Dictionary) significantly enhanced the model's capability in semantic understanding, achieving an F1 score increase from 0.817 to 0.856 on the FPB-50% dataset, an improvement of 3.9 percentage points; on FiQA Task 1, the F1 score improved from 0.820 to 0.842, a gain of 2.2 percentage points. This indicates that dictionary embeddings can effectively enhance the model's semantic grasp of financial terminology, providing supplementary information on vocabulary-level sentiment orientation.

The introduction of the neutral feature extractor (BERT-Neutral) also brought significant performance improvements, achieving an F1 score of 0.87 on the FPB-50% consensus dataset, comparable to the performance of the final EnhancedFinSentiBERT model on this dataset. This result is particularly noteworthy, suggesting that the neutral feature extractor can already provide performance close to that of the complete model when processing the FPB-50% dataset. This phenomenon may be related to the characteristics of the FPB-50% dataset, where neutral sentiment accounts for approximately 59.41%, far higher than other categories, allowing the specially designed neutral feature extractor to demonstrate significant advantages on this dataset. On the Headline dataset, BERT-Neutral achieved an F1 score of 0.932, an improvement of 3.7 percentage points over the baseline model, further confirming the critical role of neutral feature extraction in capturing subtle sentiment expressions in financial texts.

Although financial domain pre-training (BERT-financial) achieved consistent performance improvements across datasets, its gains were relatively limited, with the F1 score on the FPB-50% consensus dataset only increasing from 0.817 to 0.837, a growth of 2 percentage points. This phenomenon may be attributed to limitations in the scale of pre-training corpus and the implicit nature of domain knowledge acquisition.

Interactions between components display complex patterns. Financial pre-training combined with neutral feature extraction (BERT-financial-Neutral) outperformed financial pre-training alone on all datasets, reaching an F1 score of 0.861 on the FPB-50% dataset and 0.947 on the Headline dataset. Notably, on the FPB-50% dataset, using the neutral feature extractor alone with an F1 score of 0.870 outperformed BERT-financial-Neutral with an F1 score of 0.861, further confirming that on datasets dominated by neutral expressions, a specialized neutral feature extractor may be more effective than a combined model.

EnhancedFinSentiBERT, as the complete model, maximized overall performance by integrating all components, achieving optimal performance across all four test sets: an F1 score of 0.870 on FPB-50%, basically on par with BERT-Neutral; an F1 score of 0.980 on FPB-100%, significantly outperforming all variant models; an F1 score of 0.880 on FiQA Task 1, and an F1 score of 0.976 on Headline, both substantially leading other configurations. This indicates that while single components may excel on specific datasets, the complete model demonstrated stronger comprehensive performance and generalization ability across all datasets.

The ablation experiment results reveal three key findings: First, the neutral feature extractor contributes most significantly to model performance improvement, especially when processing financial texts containing numerous neutral expressions; second, dictionary knowledge integration provides stable performance gains, particularly in understanding professional financial terminology; finally, although financial domain pre-training contributes limitedly on its own, it provides valuable performance improvements when combined with other components, indicating synergistic effects between components. These results validate the theoretical framework proposed in Section 3.2, with component performance highly consistent with theoretical expectations: domain adaptation mechanisms create foundations for other components, knowledge enhancement mechanisms improve financial terminology understanding, and boundary optimization mechanisms optimize neutral expression identification. Performance differences across datasets also demonstrate the complementary advantages of various components under specific data distributions. These findings both prove the effectiveness of the model design and clarify how the three components collaboratively build a more robust financial sentiment analysis system.

6.3. Feature distribution comparative analysis

To systematically evaluate the representation learning capabilities of the models, we selected four models: BERT, trained BERT, trained FinBERT, and EnhancedFinSentiBERT, and analysed qualitative visualizations and quantitative metrics of feature distributions across different model variants. Fig. 8 shows the t-SNE visualization, while Table 4 provides corresponding quantitative measurements of clustering quality.

The first figure on the left shows the visualization results of the untrained BERT in sentiment representation, revealing its clear limitations in emotion differentiation. The highly mixed distribution indicates no clear boundaries between positive samples (yellow), neutral samples (green), and negative samples (purple), with sentiment categories randomly scattered throughout the feature space. The lowest inter-class distance (2.2249) quantitatively demonstrates this structural deficiency, indicating that the model is unable to distinguish between different sentiment categories in financial texts.

The visualization of the trained BERT, while showing initial clustering tendencies, also exhibits some deficiencies. Although sentiment classification begins to emerge, there is still significant overlap between categories, especially between neutral and positive samples. The higher intra-class variance (12.2692) indicates unstable feature representations, with samples of the same sentiment scattered across different regions. This suggests that domain pre-training alone is insufficient for robust sentiment classification.

The visualization of the trained FinBERT shows improved clustering effects, though issues remain. While it forms a more distinct negative sentiment region in the lower area, there is still notable confusion in the upper region where positive and neutral sentiments interweave.

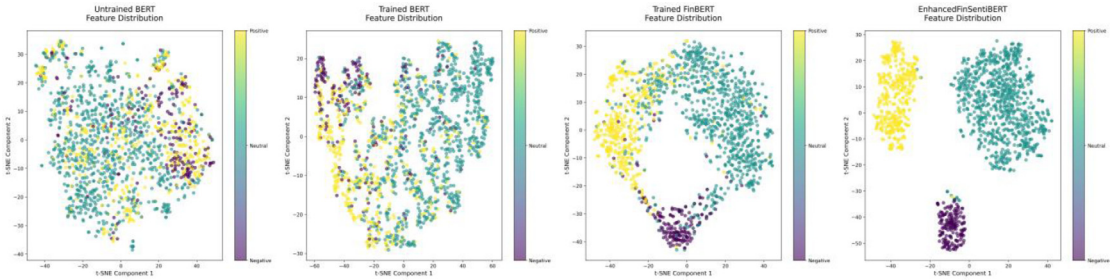

Fig. 8. t-SNE Visualization.

Table 4
Feature distribution metrics across model variants.

Model variant	Average interclass distance	Average intraclass variance
Untrained BERT	2.2249	6.6632
Trained BERT	14.0839	12.2692
Trained FinBERT	19.7771	10.2217
EnhancedFinSentiBERT	24.4092	8.4773

Although the intra-class variance has decreased, it remains considerable (10.2217), indicating persistent challenges in maintaining cluster cohesion even after extensive financial knowledge training.

The visualization of the EnhancedFinSentiBERT in the first figure on the right demonstrates the most pronounced organizational structure, forming three distinct clusters: positive sentiment concentrated in the upper left, neutral sentiment positioned in the upper right, while negative sentiment gathered into a compact cluster in the lower region. This distinct separation is quantitatively supported by the highest inter-class distance (24.4092) among all variants, while maintaining a moderate intra-class variance (8.4773).

This progression in feature distribution patterns demonstrates how our enhanced model successfully addresses the representation learning deficiencies of earlier variants. While the development of discriminative features necessarily leads to some increase in intra-class variance from the initial state, the Enhanced BERT achieves the optimal balance between cluster separation and cohesion. This comprehensive analysis provides strong evidence for the effectiveness of our integrated approach in learning robust, sentiment-aware representations of financial texts.

6.4. Case studies

In most cases, integrating lexical knowledge and neutral feature extractors improved the accuracy of financial text sentiment analysis. However, experiments also revealed occasional false predictions. To better understand the performance of the EnhancedFinSentiBERT model and the impact of lexical knowledge, this study selected representative cases, including both correct and incorrect classifications. By comparing model predictions with and without lexicon usage, the model's features and areas for improvement become clearer.

Case 1 highlights the importance of incorporating dictionary knowledge into the EnhancedFinSentiBERT model. The sentence, "The total restructuring costs are expected to be about EUR 30mn, of which EUR 13.5mn was booked in December 2008", is a typical financial reporting statement containing specific financial data and timing information. It objectively reports expected and partially booked restructuring costs without emotional bias. Using dictionary knowledge and a neutral feature extractor, the model correctly classifies the sentence as neutral, avoiding misclassification as negative. The dictionary helps the model better interpret financial terms like "restructuring costs", "expected", and "booked" as neutral. Confidence in the neutral classification increased from 44.27% to 59.24%, a 15-percentage-point improvement, demonstrating the effectiveness of this study's approach.

Without lexical knowledge, the model often misjudges such sentences as negative, likely due to general associations of terms like "restructuring costs" with negative contexts. Lexical knowledge and the neutral feature extractor enable the model to overcome this bias, accurately identifying the neutral tone typical of financial reports. This capability is critical in financial analysis, as distinguishing neutral statements from emotionally charged ones affects how investors perceive a company's financial status. Additionally, the model effectively handled specific numerical values (EUR 30mn and EUR 13.5mn), maintaining contextual accuracy despite the presence of financial data.

Case 1
sentence: The total restructuring costs are expected to be about EUR 30mn , of which EUR 13.5 mn was booked in December 2008 .
true_label: Neutral
pred_with_dict: Neutral
pred_without_dict:Negative
prob_with_dict:0.592387
prob_without_dict: 0.442718,
prob_diff: 0.149669,

However, in Case 2, the model with lexical knowledge does not show better performance than the model without lexical knowledge. While the fragment 'growing number of targeted malware attacks' may be considered negative in a general context, it is actually a neutral statement of fact in the professional context of security reporting. The dictionary word 'attacks' is labelled as a negative word, leading the model to favour a negative judgement. However, the words 'reported', 'individuals', 'companies' and 'organisations' may be neutral or even slightly positive in the financial lexicon, and their presence should have balanced out the overall sentiment prediction and made it more neutral.

Case 2
sentence: f - secure reported that : - the first half of 2008 has seen a growing number of targeted malware attacks on individuals, companies, and organizations.
true_label: Neutral
pred_with_dict: Negative
pred_without_dict: Neutral
prob_with_dict: 0.903680
prob_without_dict: 0.350213
prob_diff: 0.553467

In this case, the strong negative sentiment of the word 'attacks' can be seen as a form of noise knowledge. As a result of this noise knowledge, the financial lexicon, although relevant to financial texts in a broad sense, is less relevant and less applicable when dealing with such cross-domain technical security reports. The case reveals the inherent limitations of lexical knowledge integration in financial text analysis, especially in terms of adaptability in cross-domain scenarios.

7. Summary

This research introduces EnhancedFinSentiBERT, a financial sentiment analysis model that combines financial domain pre-training, lexical knowledge embedding, and neutral feature extraction. Validated through Financial PhraseBank, FiQA Task 1, and Headline datasets, our model outperforms baseline models such as BERT, XLNet, GPT-4, Llama, BloombergGPT, and FinBERT, especially in neutral sentiment identification. The integration of financial vocabulary knowledge enhances the model's ability to capture subtle sentiments, while the neutral feature extractor improves accuracy in handling prevalent neutral expressions. Ablation experiments indicate that although financial domain pre-training alone has limited effects, its combination with other components produces significant synergistic improvements.

While EnhancedFinSentiBERT shows promising results in financial sentiment analysis, it also has limitations. The model performs poorly on cross-domain content, particularly when integrating financial dictionaries that may introduce noise. Additionally, improvements from financial domain pre-training are not as significant as expected, possibly due to insufficient pre-training data volume and diversity.

Future research directions could focus on: (1) Expanding and optimizing the financial pre-training corpus, increasing its scale and diversity to enhance the model's underlying language understanding capabilities. (2) Improving lexical knowledge incorporation methods, exploring more flexible vocabulary integration strategies to better handle cross-domain content and reduce potential noise impacts. (3) Further optimizing the neutral feature extractor, exploring more effective extraction methods to improve its ability to identify and process subtle sentiment expressions in complex financial contexts.

CRediT authorship contribution statement

Yongyong Sun: Writing – review & editing, Writing – original draft. **Haiping Yuan:** Writing – original draft. **Fei Xu:** Supervision, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agarwal, B., 2023. Financial sentiment analysis model utilizing knowledge-base and domain-specific representation. Multimedia Tools Appl. 82 (6), 8899–8920.

Ahmad, H.O., Umar, S.U., 2023. Sentiment analysis of financial textual data using machine learning and deep learning models. Informatica 47 (5), 4673.

Cam, H., Cam, A.V., Demirel, U., Ahmed, S., 2024. Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers. Heliyon 10 (1), e23784.

Choe, J., Noh, K., Kim, N., Ahn, S., Jung, W., 2023. Exploring the impact of corpus diversity on financial pretrained language models. In: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of EMNLP 2023 2101–2112. Association for Computational Linguistics, Singapore.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Ding, X., Zhang, Y., Liu, T., Duan, J., 2015. Deep learning for event-driven stock prediction. In: International Joint Conference on Artificial Intelligence. IJCAI 2015, IJCAI, pp. 2327–2333.

Du, K., Xing, F., Cambria, E., 2023. Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis. ACM Trans. Manag. Inf. Syst. 14 (3), 23:1–23:24.

Du, K., Xing, F., Mao, R., Cambria, E., 2024. Financial sentiment analysis: Techniques and applications. ACM Comput. Surv. 56 (9), 220:1–220:42.

Garcia, D., Hu, X., Rohrer, M., 2023. The colour of finance words. J. Financ. Econ. 147 (3), 525–549.

Karanikola, A., Davrazos, G., Liapis, C.M., Kotsiantis, S., 2023. Financial sentiment analysis: Classic methods vs. deep learning models. Intell. Decis. Technol. 17 (4), 893–915.

Kaur, G., Sharma, A., 2023. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. J. Big Data 10 (1), 5.

Kearney, C., Liu, S., 2014. Textual sentiment in finance: A survey of methods and models. Int. Rev. Financ. Anal. 33, 171–185.

Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., Shah, S., 2023. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? A study on several typical tasks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, Singapore, pp. 408–422.

Lin, W., Liao, L.C., 2024. Lexicon-based prompt for financial dimensional sentiment analysis. Expert Syst. Appl. 244, 122936.

Loughran, T., McDonald, B., 2011. When is a liability not a liability?, Textual analysis, dictionaries, 10-Ks. J. Financ. 66 (1), 35–65.

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., Balahur, A., 2018. WWW'18 open challenge: Financial opinion mining and question answering. In: The Web Conference 2018 1941–1942. ACM.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P., 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. J. Assoc. Inf. Sci. Technol. 65 (4), 782–796.

Man, X., Luo, T., Lin, J., 2019. Financial sentiment analysis (FSA): A survey. In: International Conference on Power System Technology. ICPS, IEEE, pp. 617–622.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T., Trajanov, D., 2020. Evaluation of sentiment analysis in finance: From lexicons to transformers. IEEE Access 8, 131662–131682.

Sinha, A., Kedas, S., Kumar, R., Malo, P., 2022. SEntFiN 1.0: Entity-aware sentiment analysis for financial news. J. Assoc. Inf. Sci. Technol. 73 (9), 1314–1335.

Sinha, A., Khandait, T., 2021. Impact of news on the commodity market: Dataset and results. In: Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference. FICC, vol. 2, Springer, pp. 589–601.

Sohangir, S., Wang, D., Pomeranets, A., Khoshgoftaar, T.M., 2018. Big data: Deep learning for financial sentiment analysis. J. Big Data 5 (1), 3.

Wang, N., Yang, H., Wang, C.D., 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. ArXiv preprint arXiv:2310.04793.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G., 2023. Bloomberggpt: A large language model for finance. ArXiv preprint arXiv:2303.17564.

Yang, Y., Uy, M.C.S., Huang, A., 2020. FinBERT: A pretrained language model for financial communications. arXiv:2006.08097.