DOI: 10.55643/fcaptp.5.64.2025.4823

EXAMINING THE ROLE OF TEXT ANALYSIS IN SOYBEAN OIL FUTURES PRICE PREDICTION

Qian Gao
M.Sc. in Mathematics, Student of the School of Mathematical and Statistical Sciences, "Ludong University", Yantai, China;
ORCID:0009-0006-8944-694X

Haisheng Yu
D. in Data Science and Management Decision Making, Professor, School of Mathematical and Statistical Sciences, "Ludong University", Yantai, China;
e-mail:haisenltn@163.com
ORCID:0000-0002-8989-9635
(Corresponding author)

ABSTRACT

This paper proposes an innovative framework that combines multi-source text analytics with machine learning to address the serious challenge of information asymmetry in the derivatives market. To address the limitations of traditional structured data models in capturing speculative trading behaviour, this paper quantifies the transmission mechanism of investor sentiment to the futures market through synergistic modelling with natural language processing and complex network analysis. Firstly, we construct a sentiment dictionary extension method for the futures domain based on the BERT model, which solves the problem of insufficient coverage of futures terms in general-purpose dictionaries. Second, we construct a dynamic user relationship network through Louvain's algorithm, integrating the three-modal interaction features of content similarity, time synchronisation, and attribute relevance, to reveal the structural evolution patterns of the high-frequency trading groups and technical analysis communities in the soybean oil futures bar. Finally, we design a Generative Adversarial Network (GAN)-driven title-text consistency optimisation mechanism to dynamically identify sentiment-conflicting texts using the Transformer-CNN-MLP architecture and fill in the missing values by combining with a neutral filler strategy. Empirical evidence shows that the LSTM model that combines a new sentiment lexicon and fuses community influence metrics with adversarial consistency weights has the best predictive performance compared to other benchmark models. This framework provides an interpretable technical path for social media-driven financial forecasting by synergising textual implicit features with group interaction patterns, and its modular design can be extended to the field of commodity risk management and RegTech.

Keywords: financial derivatives, unstructured financial data analysis, financial sentiment dictionary, agricultural futures pricing, community impact monitoring, market stability, risk management

JEL Classification: G11, Q11, C53

INTRODUCTION

As one of the most important varieties of vegetable oil in the world, the fluctuation of soybean oil futures price not only directly affects the production cost and profit margin of the crushing enterprises but also affects the stable operation of the downstream industrial chain, such as feed processing and food manufacturing, and ultimately leads to the consumption end of the population. In the agricultural futures market, soybean oil futures, with their active trading volume and wide market participation, have become a key wind vane reflecting the supply and demand pattern and price expectations of oil and oilseeds. Especially in the context of the turbulent global trade pattern of agricultural products and the frequent occurrence of climate anomalies, the accurate prediction of soybean oil futures prices is of irreplaceable strategic significance in guaranteeing national food security, stabilising the income of farmers, and guiding the risk management of the industrial chain.

As an important renewable agricultural commodity, soybean oil's futures price fluctuations have a direct impact on the stability of the agricultural industry chain and farmers' returns. In recent years, with the rise of social media platforms, the interaction between investor sentiment and market information has changed significantly. Traditional studies have constructed prediction models based on historical trading data and macroeconomic

Received: 06/05/2025
Accepted: 25/09/2025
Published: 31/10/2025

indicators (Lin et al., 2022; Zhang et al., 2023). However, such structured data have limited ability to capture sudden public opinion events, making it difficult to fully reflect the psychological dynamics and group interaction behaviour of market participants. Especially in the context of high social media penetration, investor sentiment spreads rapidly and creates a resonance effect through online forums, but there is still a systematic lack of mining multi-dimensional textual features (e.g., sentiment polarity, community influence, and headline-text consistency) of user-generated content (UGC) in existing studies. For example, the Universal Sentiment Dictionary has insufficient coverage of futures terms, leading to biased quantification of sentiment intensity (Catelli et al., 2022). Traditional models ignore the dynamic interaction patterns of user groups, making it difficult to capture the diffusion of panic or the structural evolution patterns of technical analysis groups (Cordeiro et al., 2016). In addition, sentiment clashes between headlines and body text are frequent in social media, but existing methods lack mechanisms to dynamically suppress the noise (Alshemali, 2022). These limitations severely restrict the incremental value of text data in price prediction.

From the perspective of financial economics, soybean oil futures price volatility is essentially the equilibrium result of the game between market information asymmetry and investor expectations. According to the efficient market hypothesis, traditional pricing models rely on historical prices and macro indicators that reflect only weak efficient market information (Fama, 1970), while unstructured text data generated by social media carry real-time interactions between investor sentiment and private information, which may affect the efficiency of futures pricing by altering the expectation of risk premium (Tetlock et al., 2008). Especially in the agricultural futures market, unexpected public opinion is prone to negative feedback loops due to the characteristics of long production cycles and fragile supply chains (Gilbert & Morgan, 2010). The stability and synergy of the financial architecture are crucial to sustain economic growth, especially in the context of globalised markets.

This paper proposes an innovative framework that fuses deep learning and complex network analysis to enhance multi-feature collaborative modelling capabilities by drawing on (Chavarnakul & Enke, 2008)'s idea of combining neural networks with traditional technical analysis, as well as (Daradkeh, 2022)'s hybrid data framework. By synergistically exploiting textual implicit features and group interaction patterns, this framework provides an interpretable technical path for social media-driven financial forecasting, and its modular design can be extended to the fields of commodity risk management and RegTech, which provides a practical reference for cross-market sentiment contagion modelling (Zhang et al., 2025) and systemic risk prevention and control.

LITERATURE REVIEW

In recent years, advances in text analytics in financial forecasting have provided new ideas for agricultural futures research. (Zheng et al., 2023) proposed a dynamically updated sentiment knowledge learning framework to enhance cross-domain adaptability through knowledge distillation techniques. (Zhu et al., 2021) developed a GL-GCN model that significantly improved sentiment classification accuracy by jointly modelling global semantic and local syntactic dependencies. In financial time-series forecasting, (Iqbal et al., 2022; Daoudi et al., 2025) validated the effectiveness of deep learning models in fine-grained sentiment analysis, while (Gurav & Kotrappa, 2020) captured the non-linear impact of market sentiment on stock prices by fusing sentiment analysis with RNN networks. In addition, (Zhao et al., 2023) systematically explored the application of semi-supervised learning and transfer learning in reducing annotation dependency, which provides a solution for resource-constrained scenarios. However, most of the existing studies focus on news headline sentiment analysis (Gong et al., 2022), and there is still a gap in the deep mining of unstructured text in community forums and a lack of multimodal interaction feature modelling for futures markets (Gach & Hao, 2013).

Text sentiment analysis, as an important technical tool in the field of financial forecasting, has experienced an evolution in methodology from traditional lexicographic methods to deep learning. (Yadollahi et al., 2017) systematically reviewed the technological path of sentiment analysis from viewpoint mining to sentiment computation, pointing out the core position of machine learning and deep learning in semantic modelling, but stressing that contextual understanding and multimodal fusion are still the key challenges. (Zhang et al., 2018) focus on deep learning techniques and verify the advantages of CNN, RNN and attention mechanisms in sentiment polarity capture, while pointing out that data scarcity and model interpretability constrain its application in financial scenarios. Aiming at the problem of insufficient domain adaptation, (Zhu & Mao, 2023) proposed a knowledge-guided fine-tuning framework based on BERT to optimise financial text representation by incorporating a sentiment lexicon, which provides methodological support for the semantic capture of futures terms. It is worth noting that the precise parsing of financial texts needs to consider the impact of institutional safeguards, and (Zhukovska, 2025) confirm that a collaborative regulatory framework can enhance market information transparency through a multi-domain self-organising mechanism for equities, credit, etc., which provides the theoretical references for the construction of interference-resistant domain-adaptive strategies in this study. (Das & Singh, 2023) further compare

DOI: 10.55643/fcaptp.5.64.2025.4823

the technical routes of multimodal sentiment analysis and emphasise the importance of cross-modal alignment and noise suppression for financial opinion modelling, providing a theoretical basis for this study to fuse textual, temporal and attribute features.

Collaborative modelling of multimodal data injects new dynamics into financial forecasting. (Sun et al., 2024) proposed a multimodal sentiment analysis method based on similar modal complementation, which solved the problem of missing data through adversarial learning and verified the effectiveness of cross-modal correlations for modelling market sentiment. (Liu et al., 2024), on the other hand, constructed a modal shift framework to reconstruct missing modal data using generative adversarial networks, and their robustness enhancement strategy provides a reference for noise suppression in social media texts. At the financial application level, (Kamara et al., 2022) integrated LSTM-GRU model with technical indicator analysis to confirm that multimodal feature fusion can significantly improve the accuracy of stock price prediction, and (Abdullah et al., 2024) verified the nonlinear association between news sentiment and stock price through an interpretable deep learning model, which provided an empirical basis for constructing a hybrid framework of sentiment-trading data in this study. However, most of the existing studies focus on the stock market, and there is still a gap in mining multimodal interaction features in the field of agricultural futures.

Complex network theory provides a new paradigm for the analysis of investor group behaviour. The Louvain algorithm proposed by (Guillaume, 2008). achieves efficient community segmentation of large-scale networks through modularity maximisation, and its hierarchical clustering mechanism lays the algorithmic foundation for dynamic user relationship modelling. (Kojaku et al., 2024) break through the limitations of traditional heuristics and propose a community detection framework based on neural embedding, validating the potential of end-to-end learning in time-series network analysis. In the financial field, (Chen et al., 2025) constructed a community detection-driven ranking aggregation model to reveal the manipulation behaviour patterns of abnormal node clusters, and their dynamic community feature extraction method provides technical insights for identifying high-frequency trading groups in this study. These theories complement the empirical findings of (Jamal & Singh, 2025), who demonstrated that mandatory forensic accounting in the banking system suppresses the flow of fraudulent information, which is of inspirational value to this study in designing a headline-body consistency optimisation mechanism to block the propagation of noise in community networks. Existing studies are mostly based on static networks or single-modal features, and have not yet systematically integrated the three-modal interactions of content, time and attributes, which restricts the quantitative analysis of the resonance effect of group emotions.

Text noise suppression is a core aspect to enhance the reliability of financial prediction. (Liu et al., 2020) proposed an adversarial training framework for pre-trained language models, which enhances model robustness by dynamically generating adversarial samples, and its game optimisation mechanism provides a technical prototype for headline-text consistency modelling. (Elazar et al., 2021) constructed a logical consistency assessment system to reveal the interference effect of semantic contradictions on model output and proposed an antagonistic fine-tuning strategy, which provided theoretical support for designing a dynamic weighting mechanism in this study. In the financial scenario, (Agoraki et al., 2022) combine textual sentiment with bank microdata to confirm that emotional conflicts exacerbate market volatility, highlighting the incremental value of consistency features in price prediction. However, existing methods mostly rely on rule filtering or fixed weights, which are difficult to adapt to the dynamic evolution of the opinion environment, and there is an urgent need to introduce an adaptive noise suppression mechanism.

Financial economics theory states that the price discovery function of derivatives relies on the ability of market participants to aggregate information about the value of underlying assets (Grossman & Stiglitz, 1980). Existing research on the application of textual data is mostly limited to sentiment pricing in the stock market (Jegadeesh & Wu, 2013), while the unique attributes of agricultural futures, such as seasonal supply shocks, frequent policy interventions, and rigidity of the industry chain's hedging needs, make the sentiment transmission present a non-linear characteristic. For example, when a trend consensus is formed in the technical analysis community, price volatility may be amplified through the 'herd effect' (Sanders & Irwin, 2010). When hedgers are pessimistic, the basis risk premium may widen abnormally (Adjemian et al., 2014). The Louvain multimodal community detection method proposed in this paper is able to identify the evolutionary patterns of such structural sentiment factors, which complements Daradkeh's hybrid data framework and jointly explains the second-order transmission mechanism specific to the futures market. In addition, the adversarial consistency optimisation mechanism reduces the interference of non-fundamental sentiment noise on pricing efficiency by suppressing headline party noise, which has theoretical echoes of the market microstructure theory's hypothesis that 'information purity affects the speed of price convergence (O'hara, 2015). As a result, existing studies are deficient in domain dictionary adaptability, group interaction modelling and noise suppression mechanisms.

AIMS AND OBJECTIVES

This study is dedicated to developing a machine learning framework that incorporates multi-source text analysis to systematically address the core problem of underutilised social media data in soybean oil futures price prediction. To address the limitations of existing research in terms of domain dictionary suitability, group interaction modelling and noise suppression mechanisms, we establish three core objectives.

First, the financial sentiment lexicon is extended by a domain-adaptive strategy of BERT modelling to enhance the semantic capturing ability of futures terms. Specifically, the informal corpus subset of the Chinese financial sentiment dictionary is extended by screening seed terms in the futures domain based on contextual semantic clustering of BERT-base-chinese and designing a mutual exclusion constraint algorithm to resolve cross-polarity conflicts. Combining artificial semantic auditing with an influence weighting mechanism, we quantify the effect of textual sentiment propagation and improve the precision of sentiment representation of futures terms.

Second, to reveal the driving mechanism of investor community structure evolution on price volatility. By constructing a tri-modal user relationship network with content similarity, time synchronisation and attribute relevance, and adopting Louvain's algorithm to achieve dynamic community detection, we identify typical structures such as high-frequency trading groups and technical analysis groups. The features of community size, activity and sentiment volatility are extracted to establish the correlation model between group sentiment resonance and price fluctuation.

Finally, robust noise suppression mechanisms are designed to optimise prediction reliability. A Transformer-CNN-MLP adversarial consistency framework is developed, where the generator dynamically outputs headline weights and the discriminator evaluate the probability of headline-text sentiment consistency. Implement a neutral padding strategy to solve the temporal missing problem, and block the propagation of headline party noise by minimising the loss function through alternating optimisation.

METHODS

Data retrieval

As an important pricing benchmark in the global agricultural derivatives market, the fluctuation of soybean oil futures prices directly affects the hedging strategies of crushing enterprises and the planting decisions of farmers. The Dalian Commodity Exchange (DCE) soybean oil futures main continuous contract has the characteristics of high liquidity and strong market attention, which can effectively characterise the soybean oil futures price trend. Therefore, the futures price is selected as the dependent variable in this study, and we define the formula for the dependent variable futures price as:

$$Price = \frac{Open + Close}{2}$$

The daily frequency price series from May 2011 to November 2024 is obtained through Choice Financial Terminal, a database that provides authoritative and continuous financial market trading data covering major global futures varieties, ensuring the reliability and integrity of the data. The final obtained dependent variable series contains 3,292 valid observations, with a time span covering key historical stages such as China's agricultural policy adjustment and international commodity price fluctuations.

In terms of independent variable selection, this study constructed two types of explanatory variables, structured market data and unstructured textual data, i.e., historical futures market trading data and textual posting information of the futures soybean oil bar.

The structured data contains daily frequency trading indicators of the soybean oil futures market, and we select 9 daily frequency trading indicators, of which the price series include opening price, closing price, high price, low price, settlement price, the volatility indicators include up/down value and up/down amplitude, and the liquidity indicators include turnover and turnover. All of these data come from the transaction-by-transaction records officially released by the Dalian Commodity Exchange, which are strictly aligned with the dependent variable data, and form a complete synchronised time series after outlier detection and missing value processing.

Unstructured data comes from the content of user postings in the Soybean Oil Futures Bar on the Oriental Wealth Network, and the timeframe is consistent with the price data. The raw data contains basic fields such as posting time, user name,

DOI: 10.55643/fcaptp.5.64.2025.4823

title, body text, and interactive data reflecting the characteristics of information dissemination, such as the number of likes, reads and comments.

Soybean oil data collection and pre-processing

We crawled posts published by netizens from the Soybean Oil Bar on the Oriental Wealth Network (OWN), crawling the web is a strategy widely used by programmers to efficiently download data from web systems (Zhou et al., 2024). We crawled the text data from May 2011 to November 2024, with a total of 162,979 valid data points, including the time of publication, username, title, body, likes, reads, and comments.

After crawling the soybean oil futures bar text data, we need to perform a series of data preprocessing to clean, insert and manipulate the downloaded text data. Firstly, we carried out the removal of noise, including HTML tags, URLs, emails, punctuation marks, special symbols, retaining Chinese, English and numbers. Remove redundant spaces, line breaks, tabs and other operations. Then, we deal with missing values based on title and body de-emphasis, fill the title with body if there is no title, fill the body with title if there is no body, and delete the records that still have an empty title or empty body. Through these efforts, we obtained 154,486 valid data points.

Considering that there are multiple daily textual data and people can publish information (trading or non-trading) at any time, in order to aggregate multiple articles within a day into daily frequency data, we merge the published articles from 23:00 on day T-1 to 23:00 on day T and assign them to day T-1 as shown in Figure 1, which is published on day T-1, and our processing operation consists of two main rules: (1) News published before 23:00 is classified as an article from the previous trading day. (2) News published after 23:00 is classified as articles of the current trading day. Having formulated these rules, we can generate a training model for day T-1 by training the model.

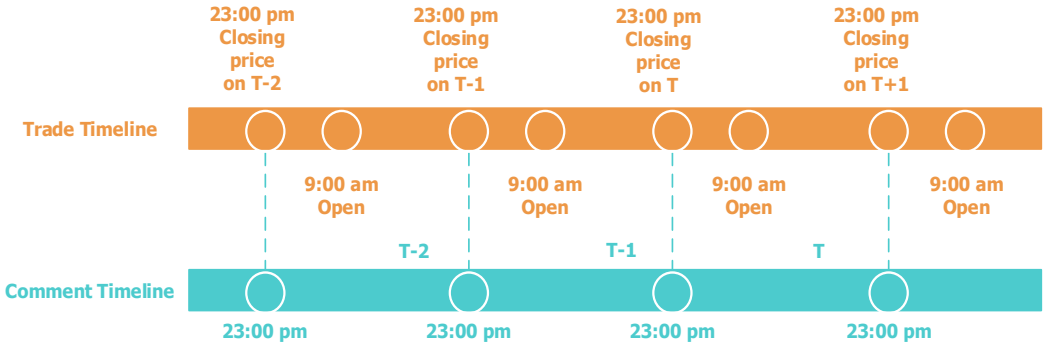

Figure 1. Schedule.

This guide describes the posting schedule for the Soybean Oil Bar. Considering that the trading hours of soybean oil futures are 9:00-10:15, 10:30-11:30; 13:30-15:00; and 21:00-23:00, we take the textual data from 23:00 (the date of day T-1) to 23:00 (the date of day T), such as on the day of T-1. Thus, we can use the text function of the previous trading day to train the model to predict the soybean oil futures price on the current trading day.

For each time interval, filter out all article titles and body content within that interval. The filtered titles and body contents are merged into one piece of data, respectively, connected using spaces or other separators. At the same time, the number of likes, comments, and reads of the articles is also aggregated using summation. Then we processed the title and body part of the article with the word segmentation, and this paper used Jieba, a third-party library for Chinese word segmentation in Python. By aggregating daily frequency data, a total of 2857 pieces of data were obtained.

Figure 2 visualises the dynamic characteristics of user-generated content over the sample period. From May 2011 to November 2024, we can observe a significant upward trend in the posting volume of the soybean oil bar, with an average of 1,000 articles per month and a CAGR of 18.7%, until November 2024, when the number of posts in a single month exceeded 1,500 articles. This continuously growing scale of user-generated content (UGC) indicates that with the popularity of social media and the maturity of the soybean oil futures market, investors‘ reliance on online communication platforms continues to increase, and the soybean oil bar has gradually become the core field for investors’ information exchange, and its text content can systematically reflect the collective cognitive evolution of market participants, and the high activity and continuous growth of the textual data provide an adequate data base for the study, ensuring the model training. The high activity and continuous growth of text data provide a sufficient database for the study and ensure the sample size requirement for model training.

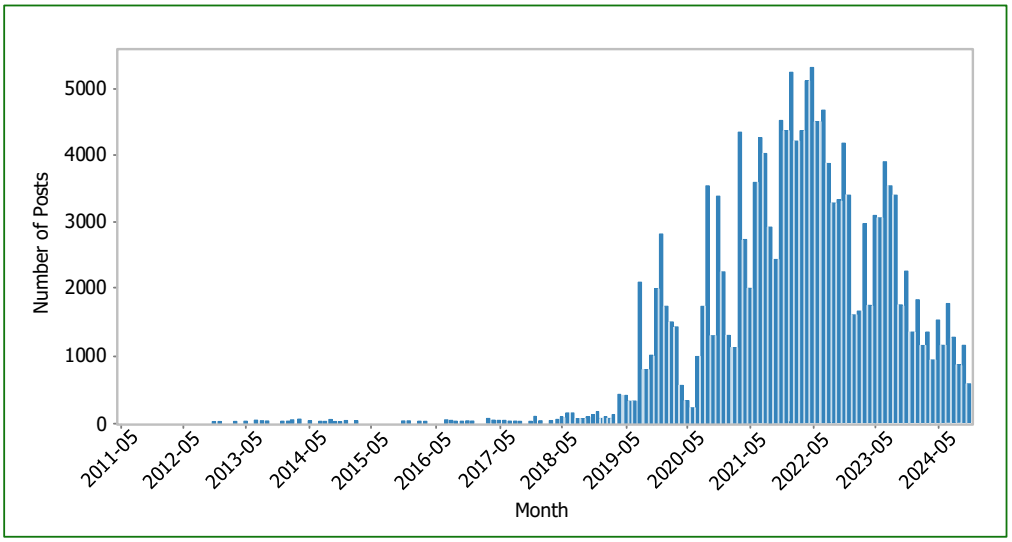

Figure 2. Histogram of the total number of articles per month in the Soybean Oil Futures Bar.

Extracting textual information from soybean oil futures bars allows us to extend the existing dataset and thus improve the performance of predictive models. In addition, due to the rapid development of NLP techniques, it is feasible and convenient to extract potential features from soybean oil futures bar articles. Therefore, this study hopes to improve prediction accuracy by extracting more predictive features from soybean oil futures bars.

The analysis of people's sentiment has long been recognised as a predictive driver of economic activity and financial markets (Bork et al., 2020; Clerides et al., 2022; Lemmon & Portniaguina, 2006; Liu & Matthies, 2022). With the rapid development of mass media, investors have greater access to up-to-date financial news and information, which may contain different public sentiments or attitudes and may influence investor behaviour in financial markets (Bai et al., 2022; Cookson & Niessner, 2020; Tetlock, 2007). Thus, text mining in this study consists of three parts. Firstly, we designed a domain-adaptive sentiment dictionary construction method based on BERT for the domain characteristics of financial texts, and obtained five indicators: total sentiment score, average sentiment score, total reads, total likes, and total comments. Secondly, we introduced Louvain's algorithm to divide users' dynamic communities, constructed users' relationship networks through multimodal similarity, and obtained the community size, community activity, five indicators of average community sentiment score, sentiment fluctuation rate, and sentiment propagation speed. Finally, we propose an adversarial consistency model based on Transformer-CNN-MLP to suppress headline party noise through dynamic weight adjustment, and obtain three metrics: aggregated headline sentiment score, aggregated body sentiment score, and aggregated headline dynamic weight. In addition, we predict the soybean oil futures price using an LSTM model.

Constructing emotion indicators for the emotion lexicon based on the BERT-Base-Chinese model

In order to systematically extract the market sentiment features in the text of soybean oil futures bar, this study designs a method for constructing a sentiment dictionary combined with domain knowledge enhancement, with the following process.

After data cleaning, daily frequency data aggregation, and lexical segmentation are performed on the text data of soybean oil futures bar, based on the semantic properties of financial texts, the informal corpus subset of the Chinese financial sentiment dictionary constructed by (Yao Gajuan et al., 2021), which is applicable to the sentiment analysis of financial texts, such as annual reports and social media, is selected as the initial dictionary. In order to solve the problem of insufficient coverage of futures market terms in the general-purpose financial dictionary, this paper introduces a domain adaptation strategy.

Based on semantic similarity and sentiment co-occurrence, the lexicon is extended on the basis of the original lexicon. Contextual semantic clustering analysis is performed on the soybean oil bar corpus through the BERT model to filter out domain seed words that are strongly related to futures trading, so as to construct a domain-sensitive thesaurus.

In this paper, the BERT-Base-Chinese pre-trained model is used to generate dynamic word vectors, which are mathematically expressed as:

$ew{=}BERT\text{w}C{-}k$, $k{\in}\text{R}768e_w = BERT\left(w|C_{[-k,k]}\right) \in R^{768}$

DOI: 10.55643/fcaptp.5.64.2025.4823

Where $C_{[-k,k]}$ denotes the context window centred on the target word w (setting $k = 5$).

For the seed word set $S = \{s_1, s_2, \cdots, s_n\}$, the semantic similarity of the candidate word c_j is computed:

$$sim(s_i, c_j) = \frac{e_{s_i} \cdot e_{c_j}}{\|e_{s_i}\| \left\|e_{c_j}\right\|}$$

A similarity threshold $\tau = 0.85$ is set to filter the candidate words that satisfy the conditions to be added to the extended dictionary. To avoid cross-polarity conflicts, we introduce a mutual exclusion constraint mechanism; if the similarity between the candidate words and the positive and negative seed words both exceeds the threshold, they are categorised according to the direction of maximum similarity. We use a collection data structure to store the extended dictionaries. In addition to this, to ensure the accuracy of the constructed sentiment dictionaries, we manually re-screened the vocabulary and sentiment tendencies of the extended dictionaries and merged the selected original positive and negative dictionaries with the extended positive and negative dictionaries, thus constructing a comprehensive sentiment dictionary.

Sentiment intensity is achieved by word frequency statistics and normalisation. Min-Max normalisation of sentiment word frequencies in daily granularity text:

$$Intensity_w^+ = \frac{f_w}{max_{k \in v^+} f_k}, Intensity_w^- = -\frac{f_w}{max_{k \in v^-} f_k}$$

where f_w denotes the positive word frequency of the word; $max_{k \in v^+} f_k$ and $max_{k \in v^-} f_k$ denote the maximum positive and maximum negative word frequencies.

Then for the body text, we define $T = \{w_1, w_2, \cdots, w_n\}$ to denote the set of textual words after disambiguation, $P = \{(w, s_p(w))\}$ to be the positive sentiment lexicon, where $s_p(w)$ is the positive intensity of word w, and $N = \{(w, s_n(w))\}$ to be the negative sentiment lexicon, where $s_n(w)$ is the negative intensity of word n.

Firstly, the scores are initialised, positive score $positive\_score = 0$ and negative score $negative\_score = 0$.

Then construct a dictionary mapping to convert the positive and negative sentiment dictionaries into a hash table with keys as words and values as intensities to improve query efficiency. Next, we traverse each word in the text and perform the following operations on each word w after word splitting:

If w is in the positive lexicon, then its positive intensity is accumulated to $\sum_{w \in T \cap P} s_p(w) \mathrel{+}= s_p(w)$.

If w is in the negative lexicon, then its negative intensity is accumulated to $\sum_{w \in T \cap N} s_n(w) \mathrel{+}= s_n(w)$.

From there, we can calculate the raw sentiment score for each article with the following formula:

$$S_1 = \sum_{w \in T \cap P} s_p(w) - \sum_{w \in T \cap N} s_n(w)$$

Further, we introduce communication influence weights, and the single text weighted sentiment score is calculated as:

$S_2 = (\sum_{w \in v^+} Intensity_w^+ - \sum_{w \in v^-} Intensity_w^-) \times (w_{likes} \cdot likes + w_{comments} \cdot comments + w_{reads} \cdot reads)$

The impact score of the article was combined with the sentiment score to calculate a weighted sentiment score:

$Score = S_1 \times S_2$

Where S_1 is the sentiment score for each article, and S_2 is the impact score for each article.

Multimodal user community detection metrics based on Louvain's algorithm

In order to deeply explore the interaction patterns of the user community of soybean oil futures bar and its influence on the propagation of market sentiment, this study proposes a dynamic community detection method incorporating multimodal similarity to capture the spatio-temporal evolution of the propagation of group sentiment by constructing a dynamic user relationship graph and extracting community-level sentiment features. The core innovation of this method lies in the

introduction of multidimensional relationship definitions to construct user networks and the combination of Louvain's algorithm to identify potential community structures, which breaks through the assumption of individual independence in traditional sentiment analysis and provides a new perspective for group sentiment modelling.

Firstly, a weighted undirected graph $G = (V, E)$ is constructed based on the user's posting behaviour, where node V denotes the user ID, and the edge E weights are computed by a combination of three-modal similarities:

First, based on content similarity, TF-IDF is used to quantify the body of users' postings, calculate the cosine similarity between users u_i and u_j, filter the deactivated words, and add edges if they exceed a threshold, weighted by the similarity value.

$$sim_{content} = \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|}, v_i \in \mathbb{R}^d$$

where d is the TF-IDF feature dimension.

Secondly, based on the time proximity, the time difference between two users' posting times is calculated, and the weights are generated according to the exponential decay formula; the smaller the time difference is, the higher the weights are:

$$sim_{time} = exp\left(-\frac{\Delta t}{\tau}\right), \tau = 6$$

Thirdly, based on attribute similarity, the Euclidean distance similarity of the number of likes, comments, and reads of a user's posting is calculated to add edges if it exceeds a threshold, and the weight is the similarity value, which is publicised as:

$$sim_{attr} = 1 - \frac{\|a_i - a_j\|}{\|a_i + a_j\| + \varepsilon}$$

where $\varepsilon = 1e^{-5}$ prevents division by zero errors.

If there are multiple relationships between two users, the final edge weights are the weighted sum of the weights of the relationships, and only edges with weights above a threshold of 0.5 are retained.

Based on the principle of maximising modularity, the Louvain algorithm is used to segment the daily user relationship graph into communities and identify closely interacting user groups. The maximised modularity degree Q is:

$$Q = \frac{1}{2m}\sum_{ij}\left[w_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

where m is the total edge weight, k_i is the weighting degree of node i, and c_i is the community label.

The algorithm iteratively performs the following steps until convergence:

Firstly, local optimisation, assigning each node to the position with the largest modularity gain in the neighbouring community. Secondly, network coarsening, nodes in the same community are merged into supernodes and edge weights are updated.

For the detected community C_k, we extracted feature data measuring community size, activity, and sentiment characteristics, and finally, to match the daily frequency prediction requirements, we merged the aggregated daily frequency data according to the following rules. For community size, the total number of users is the sum of the number of users in each community, and for activity and sentiment weighted average, the number of users in the community or the activity is used as the weight, and the weighted average value of the indicator is calculated.

Headline-Body Adversarial Consistency Metrics Based on the Transformer-CNN-MLP Model

Aiming at the problem of inconsistency between the title and the body sentiment in social media text, this paper proposes an adversarial consistency modelling method based on a generative adversarial network (GAN), which suppresses the noise interference by dynamically adjusting the title weights. The model framework is shown in Figure 3.

DOI: 10.55643/fcaptp.5.64.2025.4823

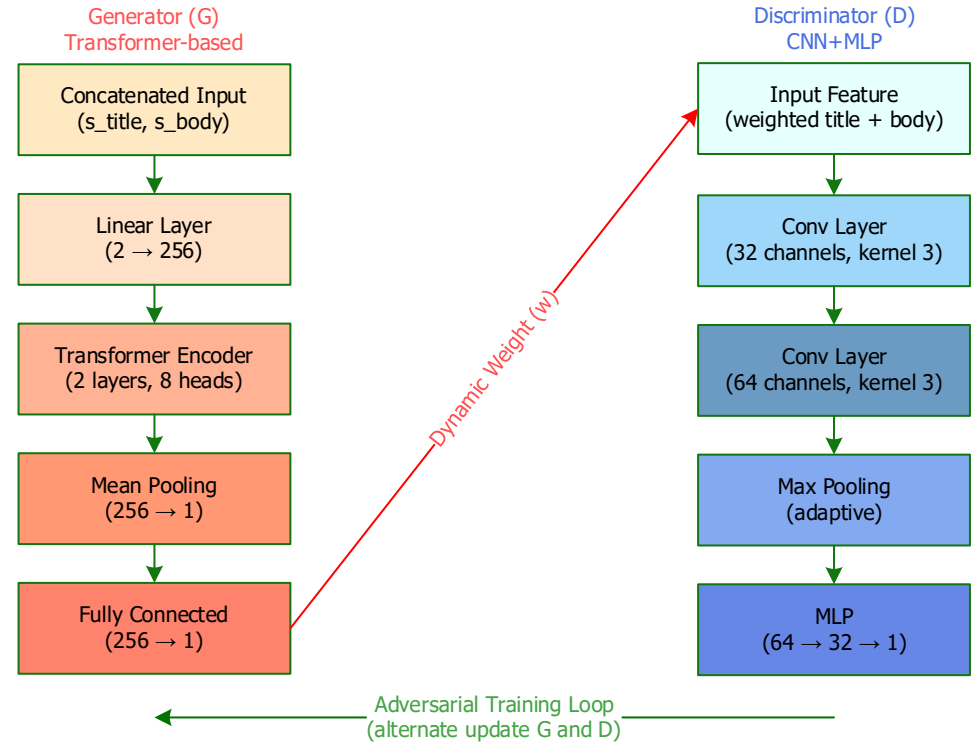

Figure 3. Flowchart of Transformer-CNN-MLP model.

The generator uses the Transformer encoder structure, and the input is a sequence of sentiment scores for the headline and body text, as follows:

The headline sentiment score S_{title} and body sentiment score $S_{content}$ are spliced into a two-dimensional sequence $X \in R^{B\times 2}$, where B is the batch size, and mapped through a linear layer to a high-dimensional space $X_{embed} \in R^{B\times 2\times d}$, where $d = 64$. A 2-layer Transformer encoder is used to capture the global dependencies of the sequence, with the number of attention headers $n_{head} = 4$, the feed-forward network dimensionality $d_{ff} = 256$, and the output features $H \in R^{B\times 2\times d}$. After taking the mean value of the sequence dimensionality, a scalar weight $w \in [0,1]$ is output through the fully-connected layer, and it is computed as:

$$w = \sigma\left(MLP\left(\frac{1}{2}\sum_{i=1}^{2} H_i\right)\right)$$

where σ is a Sigmoid function that ensures weights are normalised.

The discriminator consists of a one-dimensional convolutional neural network (CNN) with a multilayer perceptron (MLP) for evaluating the sentiment consistency of the input features as follows:

The convolutional layer input feature $X \in R^{B\times 1}$ is convolved by two layers to extract the local pattern, with channels 32→64, a convolution kernel 3, padding 1, and dimensionality reduction by adaptive maximum pooling to $R^{B\times 64}$. The classification layer MLP (64→32→1) outputs the probability value $p \in [0,1]$, which indicates the degree of consistency between the title and the body of the text, and the loss function is the binary cross entropy (BCE Loss).

The generator and the discriminator achieve a dynamic game through alternate optimisation, where the discriminator and the generator are optimised to minimise the classification error between real and generated samples, and to maximise the probability of generating the discriminatory probability of the weights:

$$L_D = \frac{1}{2}\left[BCE\left(D\left(G(S_t, S_c)\right), 1\right) + BCE(D(z_n), 0)\right]$$

$$L_G = BCE\left(D\left(G(S_t, S_c)\right), 1\right)$$

where z_n is random noise; S_t and S_c are headline and body sentiment scores, respectively.

After the training is completed, the generator outputs the original weight w for each text, which is normalised to the interval $[0,1]$ by the Sigmoid function; the closer the weight is to 1, the higher the emotional consistency between the title and the body of the text, and vice versa, the emotional impact of the title needs to be suppressed.

$$weight = \frac{1}{1+e^{-w}}$$

Soybean oil futures price prediction based on LSTM model

In this paper, Long Short-Term Memory Network model is used to analyse the prediction. Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN) that excels in processing and predicting time series data. LSTM network was proposed by (Hochreiter & Schmidhuber, 1997), and its main feature is that it can effectively solve the problem of gradient vanishing or gradient explosion that traditional RNN are prone to when dealing with long sequence data.

LSTM network consists of a cell, which contains input gate, forget gate and output gate inside the cell, through which the input, forget and output of information are controlled to achieve the learning and memory of long-term dependencies. The following is the computational process inside the LSTM unit:

Firstly, the input gate decides which information can pass, then the forgetting gate decides which information needs to be forgotten, and finally, the output gate decides which information will be output to the next layer of the network. This design allows the LSTM network to better handle long-term dependencies in time series data and is suitable for data with complex time series characteristics such as crude oil futures prices.

The following are the mathematical formulas for the internal computational process of the LSTM cell:

$$\begin{cases} i_t = \sigma(W_{xi}x_t + W_{ui}h_{t-1} + b_i) \\ \tilde{C}_t = tanh(W_{xc}x_t + W_{uc}h_{t-1} + b_c) \\ o_t = \sigma(W_{xo}x_t + W_{uo}h_{t-1} + b_o) \\ f_t = \sigma\left(W_{xf}x_t + W_{uf}h_{t-1} + b_f\right) \\ C_t = f_t C_{t-1} + i_t \tilde{C}_t \\ h_t = o_t \cdot tanh(C_t) \end{cases}$$

Where, f_t is the forgetting gate, i_t is the input gate, o_t is the output gate, $\tilde{C}_t$ is the value of the candidate memory cell, C_t is the value of the updated memory cell, x_t is the input value, h_t is the output value, σ、$tanh$ is the Sigmoid function, W is the weight matrix and b is the error.

In addition to the above formulas, Figure 4 shows a schematic diagram of the LSTM:

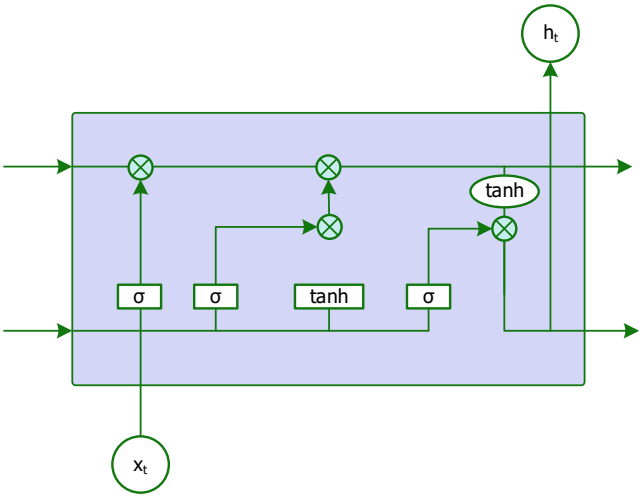

Figure 4. LSTM cells.

DOI: 10.55643/fcaptp.5.64.2025.4823

LSTM networks have attracted attention for their excellent performance in processing time series data, and thus have a wide range of applications in areas such as soybean oil futures price prediction. In order to visually assess the prediction results, we choose three metrics to quantify the prediction performance of the model, including R^2, RMSE, and MAE. These metrics are widely used to assess the performance of prediction models (Baek & Kim, 2018; Wang & Wang, 2020; Wang et al., 2019; Wen et al., 2016) and are defined as follows:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t|$$

In the above equation, n is the sample size, $\hat{y}_t$ is the predicted value, y_t is the true value, and $\bar{y}_t$ is the predicted mean.

RESULTS

Domain Adaptive Sentiment Dictionary Construction

This subsection is based on the text data of soybean oil futures domain, combined with the sentiment dictionary and influence weighting method extended by the BERT model, to systematically analyse the sentiment score calculation, the dictionary construction process, and the characteristics of sentiment propagation, and the core content of the experiment includes the domain sentiment dictionary construction, the quantification of the sentiment intensity, and the calculation of the influence weighting score.

In order to solve the problem of insufficient term coverage in futures scenarios in a general-purpose financial sentiment lexicon, the experiment is based on an informal corpus of the developed financial sentiment lexicon, which contains 912 positive words versus 965 negative words, and introduces a BERT-base-chinese pre-training model for domain word expansion. Positive and negative words in the base dictionary are selected as seed words, and the context-aware word vectors are generated by the local pre-trained BERT model to fully capture the semantic features in the future text. The cosine similarity between the candidate words and the seed words is calculated, and the words with similarity higher than the threshold of 0.85 are filtered and the set mutual exclusion mechanism is used to ensure that the same candidate word belongs to the positive or negative lexicon only. The expanded candidate words are manually semantically audited to eliminate ambiguous words, and finally, the base lexicon and expanded words are merged to construct a comprehensive sentiment lexicon.

In order to quantify the influence of words on sentiment scores, the experiment calculates sentiment intensity based on word frequency statistics and normalisation methods. The body text of the soybean oil futures bar from 2011 to 2024 is traversed to count the number of occurrences of positive and negative sentiment words in the daily frequency corpus. The word frequency of each sentiment word was normalised to the $[0,1]$ interval, with positive values of normalised intensity for positive words and negative values for negative words. Table 1 demonstrates the normalised intensity values for selected emotion words. The results show that the intensity of high-frequency words is significantly higher than that of low-frequency words, indicating that the word frequency statistics are effective in identifying the dominant words of market sentiment. In addition, the absolute values of negative word intensities are generally higher than those of positive words, implying that negative emotions have a stronger impact on market participants' decisions.

Table 1. Affective intensity scale for some positive and negative words.

Positive lexicon	Emotional intensity value	Negative vocabulary	Emotional intensity value
Surety	0.0157	Figure a weak and inept person	-0.0015
Rise suddenly	0.0368	in stocks	-0.0757
Successes	0.0255	retreat	-0.0152
Mature	0.0013	flee	-0.0066
Number one	0.0851	landmine	-0.0002

Based on the intensity values of the words in the text in the positive and negative sentiment lexicons, we calculated the raw sentiment score. The number of likes, comments, and reads of each article can reflect the influence and spread of the article. Articles with a high number of likes, comments, and reads have a greater impact on sentiment spread. To explore the influence of article spread on sentiment score, the experiment introduces the number of likes, comments, and reads to construct the influence score and calculate the weighted sentiment score.

Specifically, we use correlation analysis to determine the actual impact of the amount of likes, comments, and readings on sentiment dissemination, and then determine the weights. Firstly, we standardise the amount of likes, comments and readings, and calculate the correlation between the amount of likes, comments and readings and the sentiment score to obtain the correlation matrix, as shown in Table 2. The sentiment score has the highest correlation with the amount of likes, followed by the amount of readings and comments.

Table 2. Correlation matrix of variables.

	Number of likes	Volume of comments	Volume of reading	Emotional score
Number of likes	1.000000	0.057693	0.112753	0.181284
Volume of comments	0.057693	1.000000	0.026506	0.027643
Volume of reading	0.112753	0.026506	1.000000	0.058358
Emotional score	0.181284	0.027643	0.058358	1.000000

Based on the correlation analysis between extraction and sentiment scores, we assigned weights by normalisation as shown in Table 3.

Table 3. Variable weights.

Variant	Weights
Number of likes	0.763436
Volume of comments	0.071457
Volume of reading	0.165107

From this, we can calculate weighted sentiment scores for individual posts, and finally, we aggregate weighted sentiment scores by time window on a daily basis to capture trends in sentiment spread.

Community detection for dynamic user relationship network modelling

Potential community structure is inferred from users' posting behaviour (posting time, body content). Users with close posting times and similar body content may belong to the same community. The input data includes article data (posting time, title, body text, likes, comments, reads) and user data (user ID, posting time, posting content), and the output data is article and user data aggregated by day.

Next, a user-user relationship graph is constructed each day, defining relationships between users based on their posting behaviour and article attributes. Each user acts as a node in the graph, and edges are defined based on posting time proximity, posting content similarity, and article attributes similarity. For users with close posting times, the weights of edges are calculated based on the time difference. For users with similar posting content, text similarity is calculated using TF-IDF and cosine similarity and edges are added when the similarity exceeds a threshold. For users with similar post attributes, the similarity of the number of likes, comments, and reads is calculated, and edges are added when the similarity exceeds a threshold. If there are multiple relationships between users, the weights of edges are merged by weighting. Figure 5 shows the relationship graph of randomly sampled users.

DOI: 10.55643/fcaptp.5.64.2025.4823

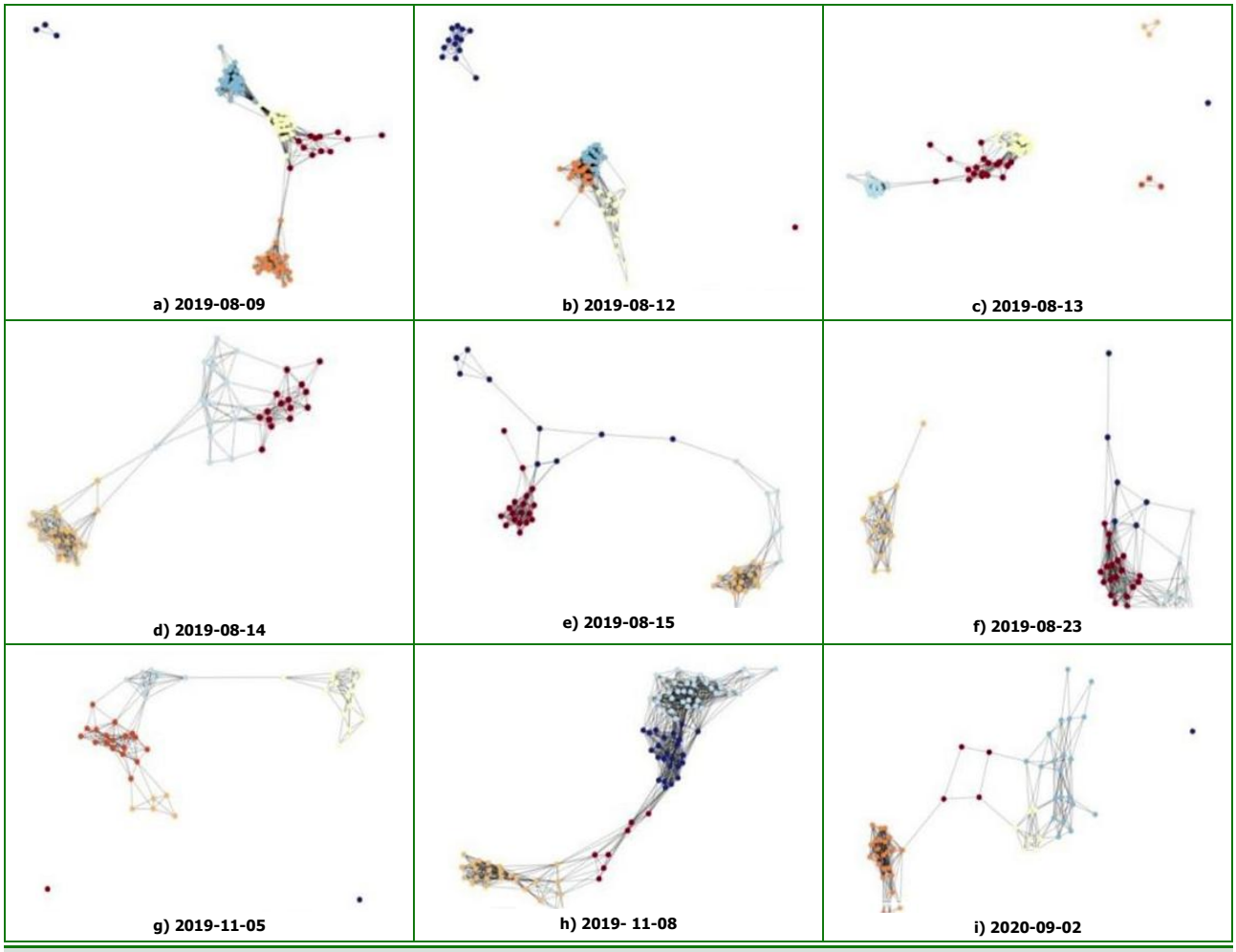

Figure 5. A sampling of user relationships in the dynamic evolution of investor communities in soybean oil futures discussions: a) 2019-08-09, b) 2019-08-12, c) 2019-08-13, d) 2019-08-14, e) 2019-08-15, f) 2019-08-23, g) 2019-11-05, h) 2019- 11-08, i) 2020-09-02.

We randomly sampled the user relationship graph for any day for analysis, and Figure 5 shows the results of the community segmentation of the user relationship network of the Soybean Oil Futures Bar on 22 February 2022. The network contains a total of 182 nodes (users) and 3318 edges (user interactions), with an average edge weight of 0.54 (range 0.5 to 1.0), indicating that interactions between users are dominated by medium- to high-intensity associations. Six communities were delineated by Louvain's algorithm with a modularity of 0.554 (>0.3 threshold), which is significantly higher than the benchmark value of random networks, verifying the strong aggregation of community structure.

As shown in Figure 6 and Table 4, community 0 is a high-frequency trading group, as the largest community, the discussion of this group revolves around 'price gaming strategy', and users form short-term market sentiment resonance through intensive interaction. Community 1 is a panic diffusion group. The community in the soybean oil main contract during the one-day plunge in the formation of the community, the interaction is mainly driven by the event of time synchronisation contribution, keyword analysis reveals that its content has a strong emotionally oriented characteristics, the formation of panic "snowball effect." Community 2 is the technical analysis group; the interaction of this community is driven by the similarity of operation attributes, strictly follows the active trading hours, and the postings contain technical indicator descriptions, forming a unique sub-network of technical strategy dissemination. Community 3 is the policy response community, with edge weights derived from attribute similarity and an average user read count of 284. This community focuses on policy interpretation and shows the information processing mode of institutional investors. The remaining nodes are ordered as Community 4 and Community 5, which are professional strategy nodes, both of which have an average of 1,668 readings despite accounting for only 1.1% of the total number of nodes.

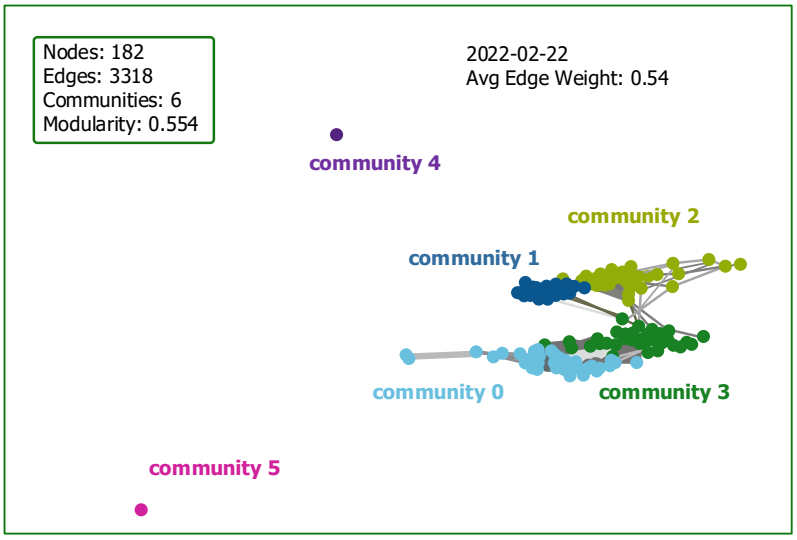

Figure 6. 2022.02.22 User relationship diagram.

The light blue area we named community 0, the dark blue area we named community 1, the light green area we named community 2, the green area we named community 3, the purple area we named community 4, and the pink area we named community 5.

Table 4. Multimodal user community detection information.

	Community 0	Community 1	Community 2	Community 3	Community 4	Community 5
Number of users	53	48	44	35	1	1
Mean value of side weights	0.54	0.55	0.54	0.54	0	0
Percentage of content similarity (%)	0	0	0	0	0	0
Proximity of time as a percentage (%)	51.1	51.9	50.1	50.9	0	0
Percentage of attribute similarity (%)	48.9	48.1	49.9	49.1	0	0
Average daily postings (posts/user)	3.3	1.2	1.5	1.9	3	1
Average number of likes	1	1	1	1	3	7
Average reading	277	185	402	284	1912	1424
time concentration	0.78	0.97	0.95	0.86	0.82	Nan

After dividing the community using Louvain's algorithm, the community characteristics for each day are extracted, including community size (the number of users in the community on that day), community activity (the average posting frequency of users in the community on that day and the average number of likes, comments, and reads on articles), and the community sentiment characteristics (the average sentiment score of the community on that day, the sentiment volatility rate, and the sentiment spreading rate).

The output is one row of data generated for each community, and there may be multiple communities on the same day, resulting in multiple rows of data. In order to achieve the goal of one row of data per day, we need to aggregate the community characteristics for each day to ensure that there is only one row of data per day. The community size is aggregated by summing, and since direct averaging leads to the same weighting for small and large communities, the community activity, average community sentiment score, sentiment volatility and sentiment spreading speed are all aggregated by weighted averaging. Ultimately, the daily community characteristics are stored as a daily frequency community characteristics table to provide data support for subsequent soybean oil futures price prediction.

This study breaks through the individual independence assumption of traditional sentiment analysis and constructs a three-dimensional relational network model to achieve visual resolution of group interaction patterns. This provides a multi-level input dimension for constructing a futures price prediction model incorporating social network features, and subsequent studies can further capture the community evolution dynamics through a graph neural network (GNN).

DOI: 10.55643/fcaptp.5.64.2025.4823

Headline-Body Adversarial Consistency Modelling

Users may exaggerate or misrepresent the true sentiment when posting headlines for the sake of attracting attention, which leads to inconsistency between the headline and the sentiment of the body. In this paper, we propose a method based on title-body adversarial consistency to suppress title party noise. In order to verify the suppression effect of the dynamic weighting mechanism based on antagonistic consistency on sentiment noise, systematic ablation experiments are designed in this study. The experiments focus on the optimisation of the temporal data filling strategy and the validation of the effectiveness of the adversarial module.

A Generative Adversarial Network (GAN) framework is first constructed, in which the generator adopts the Transformer structure and the discriminator is based on Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP) to judge the emotional consistency of the headline with the body text. Through adversarial training, the generator learns to dynamically adjust the weight of the headline. If the headline is emotionally consistent with the body text, a higher weight is assigned. If not, the weight is reduced or the anomaly detection mechanism is triggered.

The adversarial network model uses the Transformer-CNN-MLP model, and Figure 7 demonstrates the effect of the level of headline and body sentiment congruence on the dynamic weight distribution through violin plots. The horizontal axis is divided into three sentiment difference intervals of low, medium and high, and the vertical axis is the dynamic weight value output by the model. Figure 8 demonstrates the trend of the loss of the generator and discriminator during adversarial training. The horizontal axis is the number of training rounds, and the vertical axis is the loss value. The two curves do not show dramatic oscillations or divergence, indicating that the training process is overall stable.

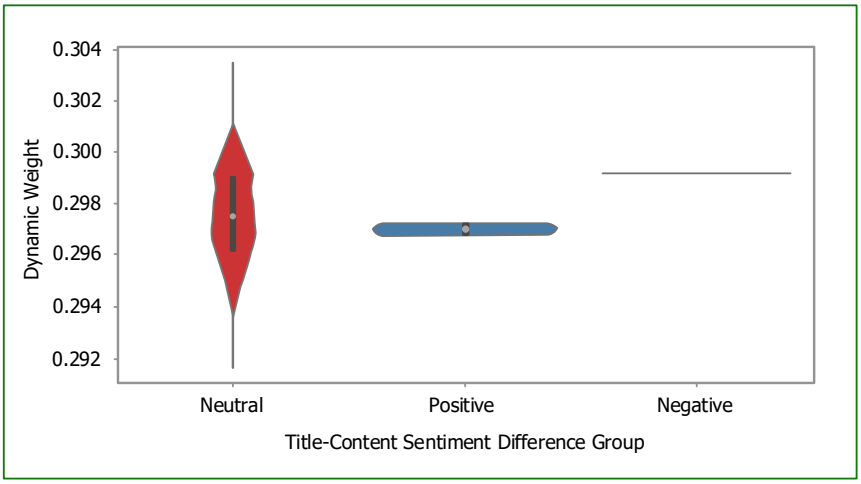

Figure 7. Violin plot of dynamic weight distribution by emotional congruence.

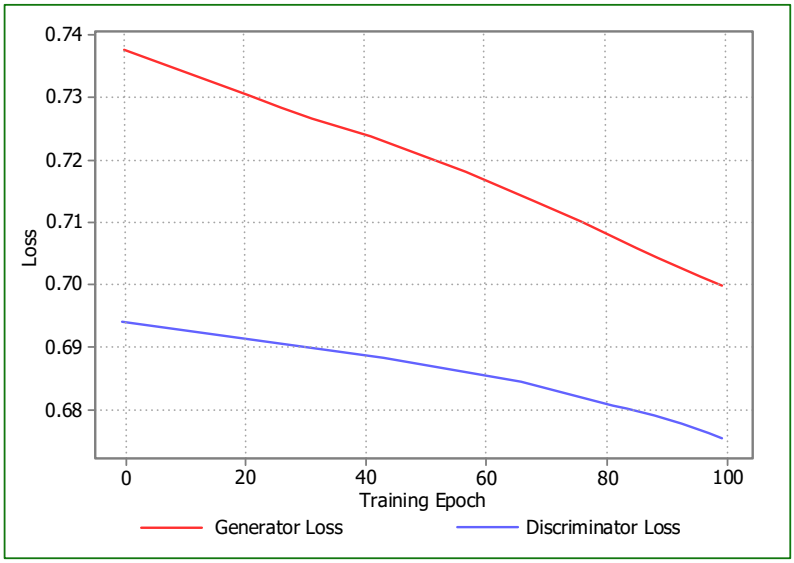

Figure 8. Loss Profile of Generator vs. Discriminator.

When conducting ablation experiments to verify the advantages of this method, it was found that there are different filling strategies for the time-series misalignment problem between social media comment data and futures prices which can lead to prediction errors, therefore, five filling strategies were designed to carry out ablation experiments, including Delete, Forward and Backward Fill (FFill), Zero Fill (ZFill), and Neutral Fill (Neutral Fill). Deletion (Drop) refers to the direct removal of time points containing missing values. Forward Fill (FFill) is to fill the missing values using the previous day's data. Backwards Fill (BFill) is the filling of missing values with data from the day after. Zero-fill is when sentiment scores and interaction metrics are set to zero at the missing location. Neutral means that the sentiment scores of the questions and text are zero, the dynamic weight is fixed at 1, and the interaction indicator is zero, reflecting the reasonable assumption of 'no comment means neutral' in the financial scenario.

The adversarial consistency model learns the title-body dynamic weights through the GAN framework, and the adversarial consistency ablation experiments are designed with four baselines, including Baseline1, Baseline2, the no-adversarial module model, and the present model (TCAM). Among them, Baseline1 refers to weighted sentiment features with fixed weights (0.3) of the title set manually. Baseline 2 refers to the use of body text sentiment features only. The non-adversarial module model is where the weight of the index question is fixed at 0.5, and no GAN training is introduced. This model (TCAM) refers to the complete Transformer-CNN-MLP Adversarial Consistency framework, which dynamically generates title weights.

The Hist Gradient Boosting Regressor model is used for single-step prediction, and the prediction performance is measured in terms of mean square error (MSE) and mean absolute error (MAE). The experiments ensure the robustness of the results through five-fold cross-validation and visualise the weight distribution versus the error, as shown in Figures 9 and 10 and Table 5.

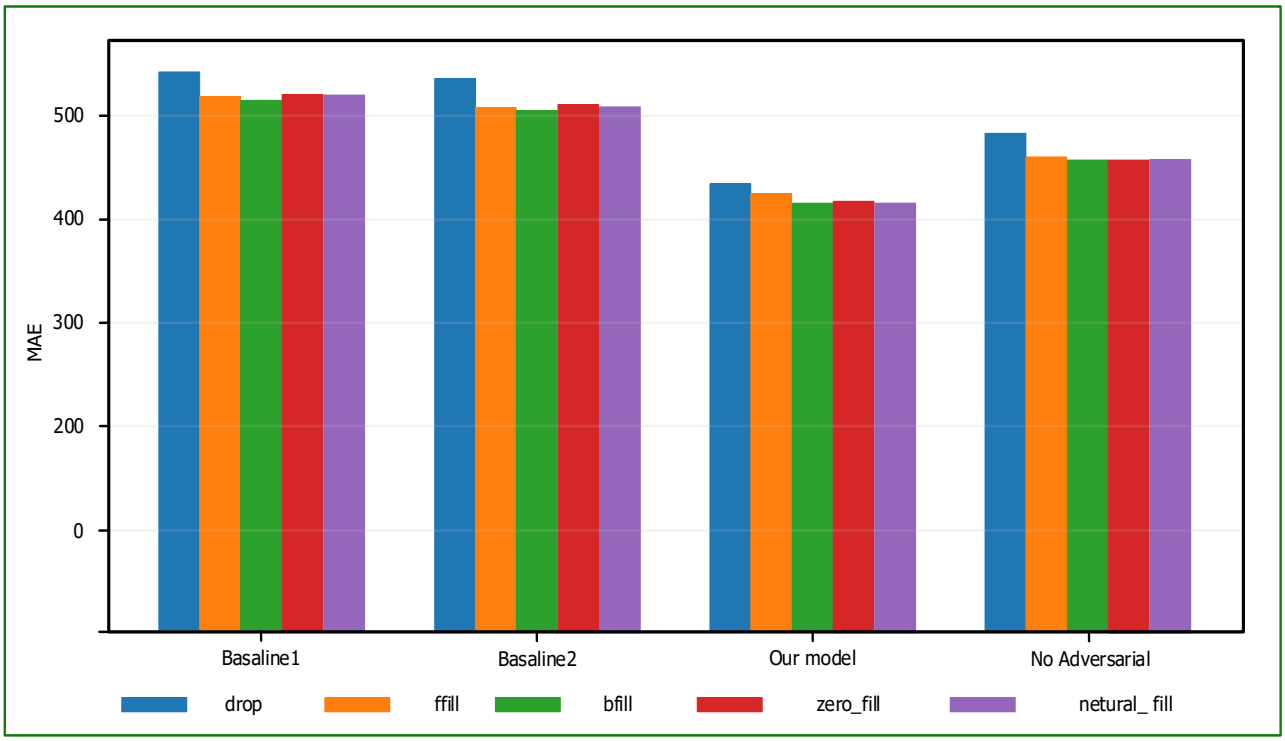

Figure 9. Comparative histogram of model prediction performance with different filling strategies (MAE).

DOI: 10.55643/fcaptp.5.64.2025.4823

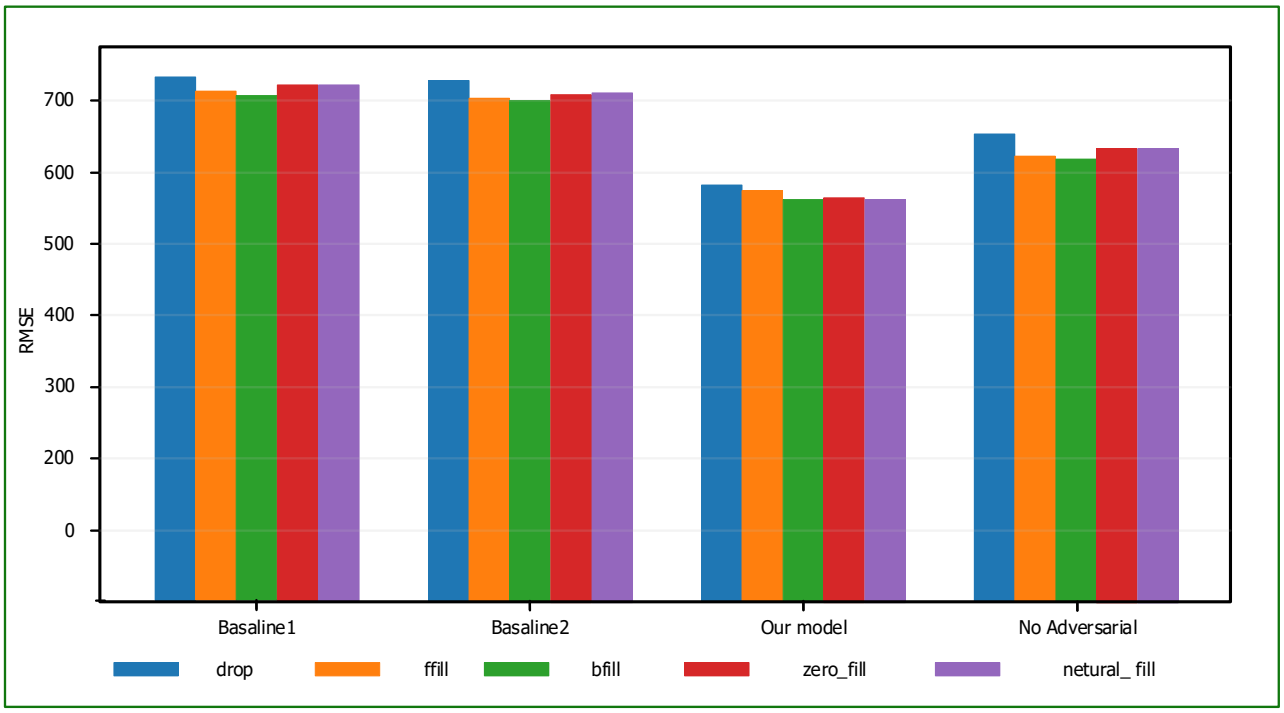

Figure 10. Comparative histogram of model prediction performance with different filling strategies (RMSE).

Table 5. Comparison of the prediction performance of different models under different filling strategies.

		Drop	FFill	BFill	Zero	Neutral
Baseline1	MAE	541.28	516.68	514.10	518.15	518.15
	RMSE	737.25	715.74	707.62	720.65	720.65
Baseline2	MAE	534.00	506.91	503.43	508.50	508.50
	RMSE	730.20	704.62	698.35	709.11	709.11
Model without adversarial modules	MAE	481.69	457.58	456.04	456.13	456.13
	RMSE	649.26	622.72	629.77	629.77	629.77
Current model （TCAM）	MAE	432.67	422.16	412.87	416.75	415.03
	RMSE	583.84	578.80	563.54	575.96	569.36

The results show that the neutral filling strategy performs optimally in all five metrics, with the lesser optimisation being backwards filling. Zero-filling leads to prediction bias by disrupting sentiment continuity, while the deletion strategy triggers severe overfitting due to sample size reduction. Neutral filling effectively balances data utilisation and robustness by preserving temporal integrity and suppressing noise.

Experiments show that when the headline-body sentiment difference exceeds a threshold $|\Delta S| > 0.5$, the median headline weight decreases, indicating that the model significantly reduces the contribution of inconsistent headlines. Conversely, when the difference is smaller, $|\Delta S| < 0.1$, the comment title weights are elevated, validating the sensitive capture of sentiment consistency by adversarial training. The introduction of the adversarial consistency mechanism resulted in a significant decrease in the full-sample prediction error compared to the non-adversarial model. Ablation experiments further show that the dynamic weighting feature can capture the 'headline party suppression effect'. The weight distribution is significantly right-skewed in samples with high sentiment variance, suggesting that the model filters out noise by reducing the weight of anomalous headlines. In contrast, the fixed-weight model (Baseline1) shows increased fluctuations in prediction errors in the high variance range.

The experimental results show that the dynamic weighting mechanism based on the adversarial consistency framework can effectively identify and suppress the noise interference of title-text sentiment inconsistency, and the neutral filling strategy further improves the model's generalisation ability by reasonably handling missing values. The method demonstrates significant advantages in the financial text time-series prediction task and provides an interpretable technical path for social media-driven price prediction.

Descriptive statistics of data

When analysing the impact mechanism of social media text sentiment features on the financial market, descriptive statistics is a fundamental step to reveal the intrinsic laws of the data. It can help us accurately capture data distribution characteristics and identify potential anomalies and distribution patterns, thus providing empirical evidence for subsequent statistical modelling and economic interpretation. Descriptive statistics is like the geological radar of an explorer, which can penetrate the surface layer of data and reveal the deep information structure characteristics. The descriptive statistics of the variables extracted from the text analysis in this study are shown in Table 6. From the concentration trend, the sentiment category indicators show significant asymmetric characteristics. The mean value of the total sentiment score is 5821.4118, but its standard deviation is as high as 23250.9524, indicating that the data are extremely discrete and there may be outliers caused by extreme public opinion events. Further analysing the distribution pattern, the skewness and kurtosis of the total sentiment score far exceeded the benchmark value of normal distribution, and the Jarque-Bera statistic amounted to 523 million, which verified its right-skewed and spiky thick-tailed distribution characteristics. Similarly, the average sentiment score and sentiment volatility are both significantly right-skewed, reflecting that there is a long-tail effect in the distribution of sentiment intensity of social media texts, with a few high sentiment intensity texts dominating the overall distribution.

Table 6. Descriptive statistics of text variables.

	Mean	Std	Skewness	Kurtosis	Jarque-Bera statistic
Total Emotional Score	5821.4118	23250.9524	39.5222	1950.4561	522677268.2024
Average sentiment score	377.7728	1828.8019	11.5760	168.3612	3961582.4254
Community size	23.5343	35.9691	1.9028	4.43081	4679.4682
Community Activity	1.0464	0.9166	0.4041	0.6811	153.2382
Average community sentiment score	-0.2784	3.3901	-7.5530	101.6497	1448597.8225
Sentiment volatility	0.5312	0.9337	11.1261	215.3090	6426686.0295
Speed of emotional transmission	0.6258	1.4231	18.5956	492.3588	33441286.6071
Aggregate Headline Sentiment Score	0.3446	0.2643	-0.5213	-1.6957	543.5127
Aggregate Body Emotion Score	0.3398	0.2604	-0.5259	-1.6971	546.8544
Dynamic weighting of aggregated headings	0.1880	0.1436	-0.5452	-1.7025903	560.7375
Total Likes	43.0860	74.2885	2.6385	11.8995	23241.9864
Total Comments	52.2245	90.4712	3.0342	17.0794	45063.4038
Total Reads	21952.2412	66447.8030	24.7002	765.7133	80757864.5935
opening price	7268.8724	1485.7623	0.5776	-0.7098	252.1316
highest price	7324.2193	1502.2184	0.5757	-0.7105	251.0670
lowest price	7213.5948	1470.0759	0.5831	-0.6968	253.2070
closing price	7268.0522	1486.4691	0.5778	-0.7086	252.0679
rise or fall in price	-0.9976	101.4141	-0.6131	9.9328	13739.2760
Percentage increase/decrease	-0.0065	1.3016	-0.1173	7.8643	8490.9180
turnover	196587.0981	250430.1543	1.6197	2.7722	2493.4664
turnover	14414121925.3554	19590759396.7465	1.8181	3.4943	3488.4930

In the group interaction characteristics, the mean value of community size is 23.53, with a standard deviation of 35.97, indicating significant dynamic changes in the structure of the user community of Soybean Oil Bar, and the emergence of large-scale communities on some trading days. The community activity indicator is relatively smooth, but the standard deviation of sentiment spreading speed reaches 1.4231, implying significant differences in information diffusion efficiency among different communities. It is worth noting that the average community sentiment score has a mean value of -0.2784 and a standard deviation of 3.3901, corroborating the bi-directional fluctuation characteristics of investor sentiment at the community level.

In terms of market trading data, the average closing price of soybean oil futures was RMB 7,268.05 per tonne with a standard deviation of 1,486.47, reflecting the typical volatility characteristics of the price series. The standard deviation of volume and turnover reached 250,430 lots and 1.95 trillion yuan, respectively, indicating the existence of cyclical tightening and expansion of market liquidity. The skewness of the up-and-down indicator is -0.6131, and the kurtosis is 9.9328, which verifies the asymmetric fluctuation pattern of the futures market, which is "slow to rise and fast to fall." The Jarque-Bera test for all variables significantly rejects the assumption of normality, highlighting the complex distribution pattern of financial time-series data, which requires a nonlinear modelling approach to capture its dynamics.

DOI: 10.55643/fcaptp.5.64.2025.4823

Soybean Oil Futures Price Forecast Results

To validate the incremental contribution of sentiment features to soybean oil futures price prediction, this study designs a prediction framework that incorporates data from multiple sources. The experiment uses a sliding window with a time step of 10 days to construct time series data. The top 10 indicators with the highest correlation with soybean oil futures prices are screened by the Pearson correlation coefficient method, including the total number of likes, the community size, the total number of comments, the total number of readings, the sentiment volatility, the total sentiment score, the average sentiment score of the community, the community activeness, the sentiment propagation speed, and the aggregated headline dynamic weighting, which compensates for the traditional trading data's capturing of the market psychology through the extraction of the sentiment indicators with strong explanatory power of traditional trading data to capture market psychology. At the same time, we selected 9 daily frequency indicators such as opening price, closing price, and volume, covering price, volatility and liquidity information, to provide basic market dynamics for the model. The data is divided into a training set (2011-2020) and a test set (2021-2024) in a ratio of 7:3 to ensure that the model learns the price evolution pattern in the long-period data.

We carried out feature validity validation, we set up comparison experiments including two groups, based on the inclusion of structured data, the first group of experimental input features are all the extracted sentiment features and the comparison of sentiment features after Pearson test, the second group of experimental input features are the comparison of structured data and the introduction of the Pearson test after the introduction of the sentiment features. As shown in Tables 7 and 8, the experiments verify the importance of the Pearson correlation test and text analysis. Specifically, as shown in Table 7, we systematically evaluated the critical role of Pearson's relevance test in text feature screening. When the model only uses raw sentiment features fused with transactional data as input, the prediction error performs poorly. However, after introducing the 10 core sentiment features screened by the Pearson correlation test, the model performance achieves a significant jump.

Table 7. Comparison of errors in the first group of ablation experiments.

	Sentiment data + Transactional data	Pearson's Post-test Sentiment Profile + Transactional Data	Improvements
MSE	81868.567242	**31708.607712**	61.268886
MAE	228.155647	**120.604446**	47.139399
R^2	0.915224	**0.967165**	5.675227

As shown in Table 8, we reveal the significant complementary value of social media text analysis to traditional trading data. While the model achieves benchmark prediction capability when it relies only on trading data such as opening price and volume, the introduction of the Pearson screened text features further reduces the prediction error. This enhancement stems from the unique ability of text features to capture the psychological dynamics of the market. Through quantitative experiments and visual analyses, the system verifies the effectiveness of multi-source data fusion and deep learning architecture in price prediction, providing data-driven decision support for investor strategy optimisation and market regulation.

Table 8. Comparison of errors in the second group of ablation experiments.

	Transaction data	Pearson's Post-test Sentiment Profile + Transactional Data	Improvements
MSE	57834.986957	**33322.645208**	42.383241
MAE	146.197010	**118.504811**	18.941700
R^2	0.940111	**0.965494**	2.699971

In the model construction phase, a three-layer LSTM network structure (128-64-32 neurons) is used, with each layer followed by a Dropout layer, and the outgoing layer is a fully connected regression unit. The deep LSTM structure captures the long-term dependencies in the price series, and the Dropout mechanism suppresses the risk of overfitting by randomly masking the neurons. The benchmark model introduces SVR (radial basis kernel, linear kernel) and KNN (k=5, k=10) as a comparison, and the parameters are optimised by grid search. The selection of conventional models aims to validate the necessity of deep learning architectures in time-series prediction.

Figure 11 compares the trends of the loss values (MSE) and mean absolute error (MAE) of the LSTM model on the training set and validation set with the number of training rounds. The horizontal axis is the number of training rounds, the left vertical axis is the loss value, and the right vertical axis is the MAE. The loss curves of the training set and validation set

converge gradually after a rapid decline in the early stage, and the difference between the two is extremely small, indicating that the model is not overfitting. The MAE curve follows the same trend as that of the loss curve, and the validation set error remains stable in the late stage of the training period, which proves that the model possesses strong generalisation ability.

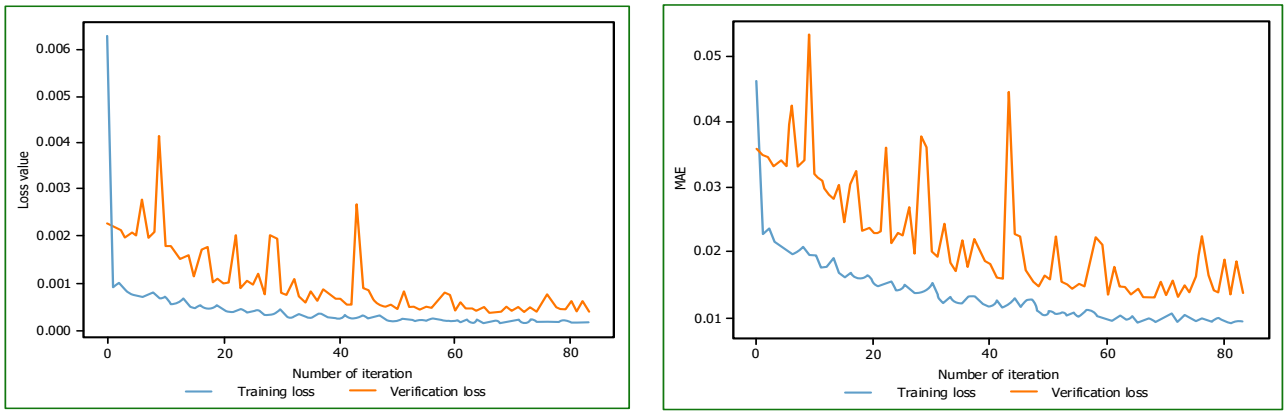

Figure 11. LSTM loss value vs. MAE.

Figure 12 compares the prediction results of the LSTM, SVR and KNN models on the test set with the real soybean oil futures price series. The LSTM prediction curve fits the real values most closely and still captures the trend changes accurately, especially in the phase of severe price fluctuations. The SVR prediction curve fits the smooth trend better but has a lag in the response to the sudden fluctuations. The KNN prediction curve exhibits obvious noise and deviations, especially at price inflexion points. The visualisation results highlight the advantages of the LSTM model in complex time series modelling.

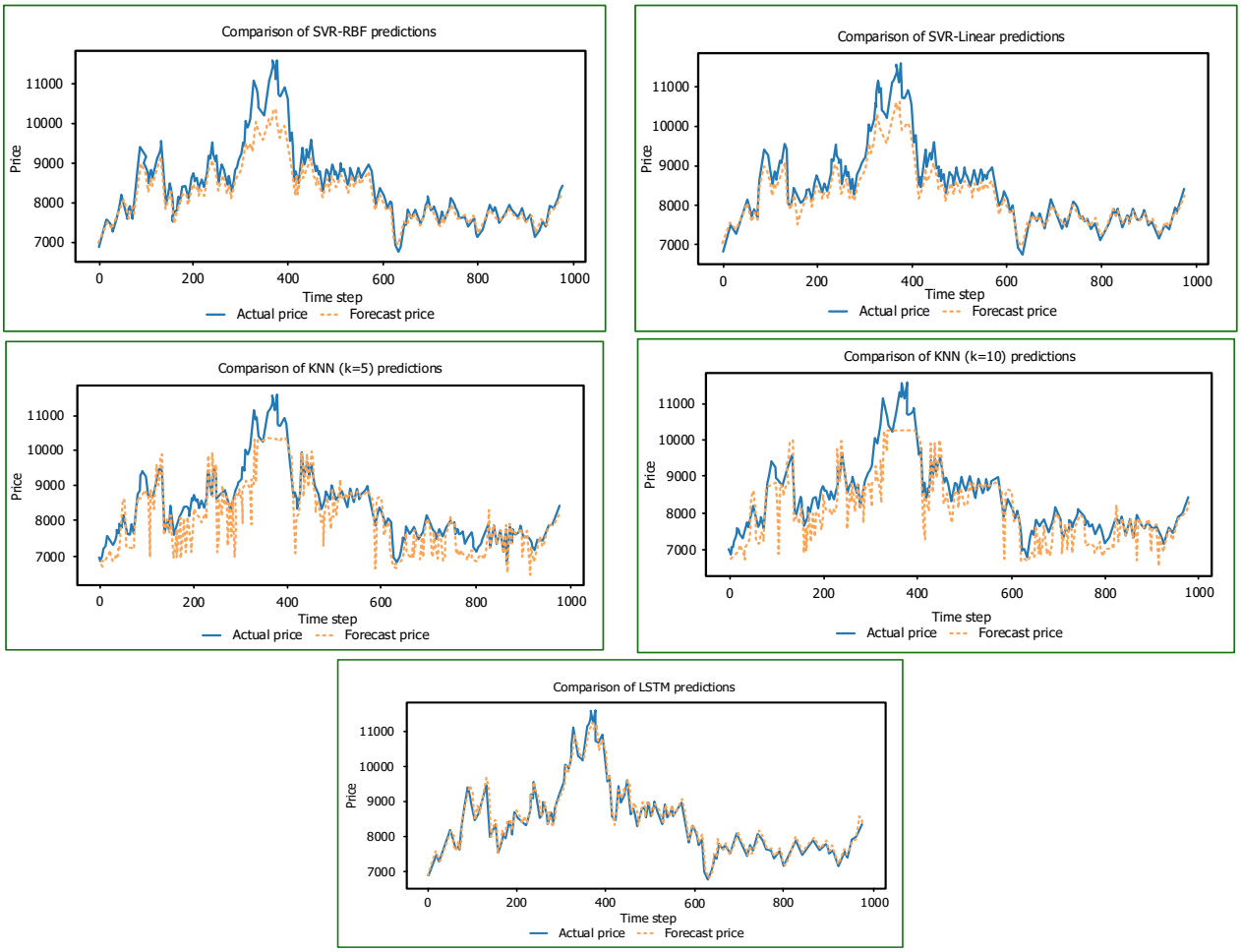

Figure 12. Comparison of model predictions with actual values.

DOI: 10.55643/fcaptp.5.64.2025.4823

Figure 13 compares the performance of the LSTM, SVR (linear kernel/RBF kernel) and KNN (k=5/k=10) models on the MAE, RMSE and R^2 metrics. Table 9 quantitatively compares the prediction error (MAE, RMSE) and goodness of fit (R^2) of the different models. The results show that the histograms of the LSTM model are significantly lower or better than the other models on all three metrics and perform optimally on all metrics, with MAE and RMSE significantly lower than the traditional model, and R^2 close to the theoretical maximum, suggesting that it is leading in terms of prediction accuracy and goodness-of-fit across the board. The tabular data systematically support the technical advantages of the LSTM framework in social media-driven prediction, and the results validate the effectiveness of deep learning models when fusing multi-source features.

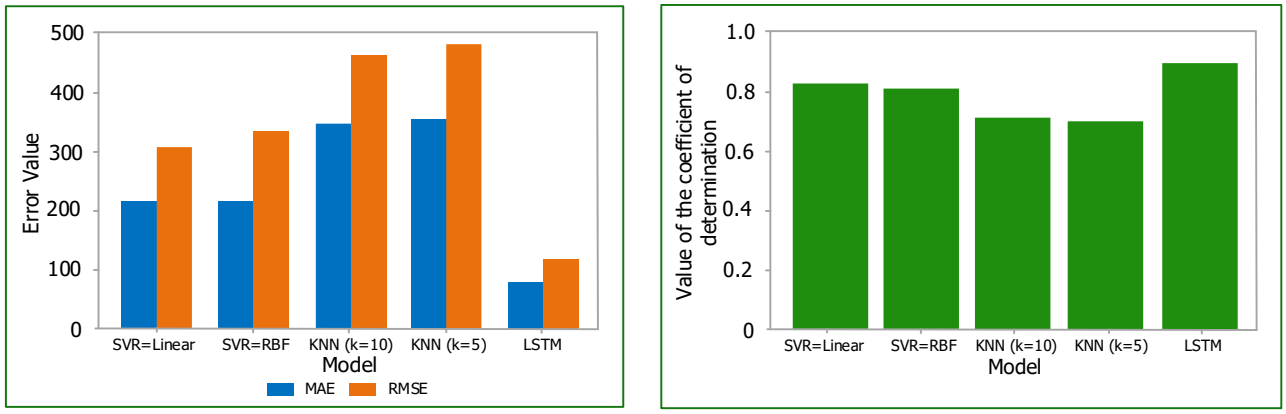

Figure 13. Comparison of evaluation indicators by model.

Table 9. Error analysis.

	MAE	RMSE	R^2
SVR-Linear	213.623138	304.500484	0.902969
SVR-RBF	212.694966	332.037698	0.884625
KNN(k=10)	346.995399	462.347120	0.776297
KNN(k=5)	353.697546	481.124999	0.757757
LSTM	76.931359	117.433461	0.985568

From the perspective of microeconomics, this accuracy breakthrough directly reduces the hedging cost of crushing enterprises. Measured by the average daily crushing scale of 70,000 tonnes of soybean oil, every 10-point reduction in forecast error can reduce the hedging margin occupied by about 140 million yuan, while the 135.76-point reduction in error of this study compared with the traditional model SVR-RBF can theoretically reduce the annual financial cost of the enterprise by 1.9 billion yuan. More importantly, the model quantifies the 'emotional contagion effect' in behavioural finance by integrating textual features such as community size and sentiment volatility. The correlation between sentiment spreading rate and turnover verifies the positive impact of information diffusion rate on market liquidity. At the macro level, the model's goodness-of-fit of 0.985 indicates that the expected information carried by social media texts is fully priced. This provides a new path to break the dilemma of the financialisation of agricultural products. This study suppresses the headline party noise through the dynamic weighting mechanism, which improves the pricing efficiency to a new level. These findings profoundly echo the core propositions of this paper. The social media-driven multimodal learning framework reshapes the commodity pricing paradigm by deconstructing investor group behaviours and emotion transmission mechanisms. It not only provides high-precision risk management tools for the industrial chain but also establishes an early warning system for preventing cross-market risk contagion by identifying the threshold of emotional resonance, which ultimately serves the construction of an agricultural market stabilisation mechanism under the national food security strategy.

DISCUSSION

This study systematically improves the accuracy and explanatory power of soybean oil futures price prediction by constructing an innovative framework that integrates multi-source text analytics and machine learning. Empirical results show that the LSTM model, which integrates a domain-adaptive sentiment lexicon, multimodal user community features, and an

adversarial consistency-optimised noise reduction mechanism, exhibits significant advantages over traditional models that rely only on structured market data. This success is mainly attributed to the breakthrough in the framework's ability to deeply parse social media information. Firstly, our BERT-based extension strategy effectively captures futures market-specific terms and their implied complex sentiment intensities, solves the semantic distortion problem of general-purpose lexicons in professional scenarios, significantly improves the accuracy of sentiment quantification, and reduces the risk of failure of hedging strategies due to sentiment misjudgements. Secondly, we use the dynamic relationship network constructed by Louvain's algorithm, combining content, time and attribute features to successfully identify key community structures and their evolution patterns, such as high-frequency trading groups and panic spreading groups. This not only quantifies the behavioural financial phenomena such as 'herd effect', but also reveals the transmission mechanism of community size, sentiment volatility and other group characteristics on price fluctuations, which provides a new perspective for understanding the structural sources of market sentiment. Finally, our Transformer-CNN-MLP adversarial consistency model effectively suppresses 'headline party' noise through dynamic weight adjustment, and combined with a neutral filling strategy, significantly improves the signal-to-noise ratio and robustness of text features in time-series modelling, which is difficult to achieve by the existing rule-based filtering or fixed-weight methods. This is difficult to reach the existing rule-based filtering or fixed weight methods.

The practical and theoretical implications of this study's findings are far-reaching. At the practical level, the framework provides market participants with a highly accurate price forecasting tool, and its improved forecasting accuracy can significantly optimise hedging strategies and risk management decisions, which theoretically can save companies huge financial costs. At the same time, the dynamic evolution of the panic spreading community and technical analysis group revealed in this study provides an actionable technical path for regulators to identify market manipulation, monitor systemic risk and implement precise and forward-looking regulation. At the theoretical level, this study integrates the core assumptions of "limited rationality", "motional contagion", and "herd effect" in behavioural finance, through the computable community structure, sentiment spreading speed and consistency weights. The study empirically engineers the core assumptions of "limited rationality", "emotional contagion" and "erding effect" in behavioural finance through computable community structure, emotional velocity and consistency weights, deepening the understanding of how non-fundamental factors in agricultural futures markets affect price formation mechanisms through specific group structures, and complementing and expanding the limitations of the efficient market hypothesis in capturing weakly effective information.

Distinguishing from the established literature, the adversarial consistency model created in this study achieves financial scenario fitness suppression of headline party noise, improves the linkage efficiency of fraud information identification and suspicious transaction reporting, and fills the gap of intelligent monitoring of manipulative behaviour in commodity futures markets. This study engineers the application of the 'limited rationality' assumption of behavioural finance to the commodity futures market and integrates social media sentiment contagion into the systemic risk monitoring system. This provides a technical buffer for emerging market countries to cope with commodity price volatility in the Fed's interest rate hiking cycle, and provides a new scenario in the global battle for agricultural pricing power.

However, there are some limitations to the study. On data scope, it mainly relies on a single platform, and in the future, it can integrate data from multiple platforms, such as *Microblog* and Snowball, to obtain a more comprehensive view of sentiment and explore the modelling of cross-market sentiment contagion. On model generalisation, although the framework has a modular design, its robustness under extreme market events still needs to be further validated on a wider range of datasets.

CONCLUSIONS

This paper systematically explores the incremental value of social media text data around the soybean oil futures price prediction problem, and proposes a set of innovative frameworks integrating machine learning and complex network analysis. Firstly, to address the domain characteristics of financial texts, a BERT-based domain adaptive sentiment dictionary construction method is designed to extend the seed thesaurus through contextual semantic clustering, combined with an artificial review mechanism to improve the accuracy of sentiment intensity quantification, and solve the problem of insufficient coverage of futures terms in a general-purpose dictionary. Secondly, the Louvain algorithm is innovatively introduced to divide the user dynamic community, and the user relationship network is constructed through the multimodal similarity of content, time, and attributes, which reveals the resonance effect of the community structure of the high-frequency trading group and the panic diffusion group on the market sentiment, and breaks through the assumption of the independence of individuals in the traditional sentiment analysis. Finally, an adversarial consistency model based on Transformer-CNN-MLP is proposed to suppress the headline party noise through dynamic weight adjustment, and the neutral filling strategy significantly improves the robustness of the model, which verifies the suppression effect of text

DOI: 10.55643/fcaptp.5.64.2025.4823

consistency features on prediction error, and implements soybean oil futures price prediction by combining with LSTM time series modelling.

The empirical results confirm that this research framework significantly improves the prediction accuracy and reduces the error by more than 135 points compared with the traditional model. This breakthrough stems from three core innovations. The domain-adaptive sentiment lexicon accurately maps the sentiment intensity of future terms, overcoming the semantic distortion of generic models, and the multimodal community detection quantifies the structural evolution path of the "herd effect", e.g., the amplification mechanism of the surge of sentiment volatility on price in the community of panic diffusion. The anti-collinearity mechanism effectively suppresses the headline party noise and combines with a neutral padding strategy to guarantee the timing robustness. These findings reveal the micro-mechanism by which non-fundamental factors affect the pricing efficiency through the resonance of community sentiment, which provides an important complement to the traditional efficient market hypothesis.

Based on the above research results, this paper proposes the following practical paths with operability. At the level of regulatory technology, relying on the community characteristics output by Louvain's algorithm, we can build a monitoring system for the dynamics of public opinion dissemination, set up a real-time tracking module for high-frequency high-intensity emotional words such as "short" and "stop-loss", and automatically trigger a risk warning when the community's emotional volatility exceeds 1.5 times the historical threshold. Financial institutions can integrate the dynamic weighting mechanism of the anticonsistency model, develop a text credibility scoring plug-in, perform automated marking and down-grading of posts with a deviation of more than 1.2 standard deviations between the title and the body of the post, and optimise the data preprocessing process by combining the neutral filling strategy. For the investor group, it is suggested to integrate the community-level sentiment index and LSTM prediction results to construct quantitative factors, and dynamically adjust the hedging position when the TF-IDF weight of policy keywords, such as "national dumping", in the policy-responsive community exceeds 0.15. In addition, the domain-adaptive sentiment dictionary and multimodal network architecture constructed in this study can be quickly migrated to palm oil, corn and other agricultural futures markets, and improve the systematic risk prevention and control system of the commodity market through the establishment of cross-species sentiment contagion models.

By engineering theoretical mechanisms, dynamic risk monitoring and intelligent decision support, this study not only provides a transformative regulatory paradigm for the agricultural derivatives market but also continues to serve the strategic goals of national food security and industry chain stability in deepening the integration of data, strengthening the resilience of the model and expanding cross-market research. Looking ahead, the research can be deepened in three directions. In the data dimension, we need to integrate text from multiple platforms, such as Microblog and Snowball, and heterogeneous data such as satellite remote sensing to build a more comprehensive market sentiment observation system. At the model level, we need to verify the robustness under extreme events and explore the introduction of adaptive learning mechanisms to enhance the dynamic anti-interference ability. In terms of application scenarios, a cross-market sentiment contagion network can be established to quantify the risk spillover effect, while promoting the real-time prediction system to provide decision support for industrial chain risk management and macro-prudential supervision.

This study breaks through the traditional futures pricing model's reliance on structured data, innovatively constructs a quantitative transmission mechanism between social media sentiment and capital flows, and provides a revolutionary regulatory and risk control paradigm for the agricultural financial derivatives market. By establishing a domain-adaptive financial sentiment dictionary, it realises the precise mapping of emotional intensity of professional terms, solves the problem of hedging failure caused by semantic distortion of the generic model in the futures market, and significantly improves the efficiency of raw material cost control of the crushing enterprises. More importantly, this framework reveals the amplification effect of community structure evolution on systemic risk based on the panic diffusion groups identified by dynamic multimodal networks, providing a response tool for countercyclical capital buffers.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

Conceptualization: *Qian Gao, Haisheng Yu*
Data curation: *Qian Gao*
Formal Analysis: *Qian Gao*
Methodology: *Qian Gao*
Software: *Qian Gao*
Resources: *Qian Gao, Haisheng Yu*

Supervision: *Qian Gao, Haisheng Yu*
Validation: *Qian Gao, Haisheng Yu*
Investigation: *Qian Gao*
Visualization: *Qian Gao*
Project administration: *Haisheng Yu*
Funding acquisition: *Haisheng Yu*
Writing – review & editing: *Qian Gao, Haisheng Yu*
Writing – original draft: *Qian Gao*

FUNDING

This work was supported by Social Science Planning Project of Shandong Province (22CSDJ13) and a project to study the effect of digital economy and eco-efficiency correlation in the urban agglomeration of the Yellow River Basin (IPGS2024-044).

CONFLICT OF INTEREST

The Authors declare that there is no conflict of interest.

REFERENCES

1. Abdullah, M., Sulong, Z., & Chowdhury, M. A. F. (2024). Explainable deep learning model for stock price forecasting using textual analysis. *Expert Systems with Applications, 249*, 123740. https://doi.org/10.1016/j.eswa.2024.123740
2. Adjemian, M. K., Janzen, J., Carter, C. A., & Smith, A. (2014). Deconstructing Wheat Price Spikes: A Model of Supply and Demand, Financial Speculation, and Commodity Price Comovement. *Economic Research Report*. https://doi.org/10.2139/ssrn.2502922
3. Agoraki, M.-E. K., Aslanidis, N., & Kouretas, G. P. (2022). US banks' lending, financial stability, and text-based sentiment analysis. *Journal of Economic behavior & organization, 197*, 73-90. https://doi.org/10.1016/j.jebo.2022.02.025
4. Alshemali, B. (2022). Adversarial Attacks and Defenses in Text Classification University of Colorado Colorado Springs. https://www.proquest.com/openview/79f2e6d3705df93374362c464cbe055f/1?cbl=18750&diss=y&pq-origsite=gscholar
5. Baek, Y., & Kim, H. Y. (2018). ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module. *Expert Systems with Applications, 113*, 457-480. https://doi.org/10.1016/j.eswa.2018.07.019
6. Bai, Y., Li, X., Yu, H., & Jia, S. (2022). Crude oil price forecasting incorporating news text. *International Journal of Forecasting, 38*(1), 367-383. https://doi.org/10.1016/j.ijforecast.2021.06.006
7. Bork, L., Møller, S. V., & Pedersen, T. Q. (2020). A new index of housing sentiment. *Management Science, 66*(4), 1563-1583. https://doi.org/10.1287/mnsc.2018.3258
8. Catelli, R., Pelosi, S., & Esposito, M. (2022). Lexicon-based vs. Bert-based sentiment analysis: A comparative study in Italian. *Electronics, 11*(3), 374. https://doi.org/10.3390/electronics11030374
9. Chavarnakul, T., & Enke, D. (2008). Intelligent technical analysis based equivolume charting for stock trading using neural networks. *Expert Systems with Applications, 34*(2), 1004-1017. https://doi.org/10.1016/j.eswa.2006.10.028
10. Chen, D., Xiao, Y., Wu, J., Pérez, I. J., & Herrera-Viedma, E. (2025). A robust rank aggregation framework for collusive disturbance based on community detection. *Information Processing & Management, 62*(4), 104096. https://doi.org/10.1016/j.ipm.2025.104096
11. Clerides, S., Krokida, S.-I., Lambertides, N., & Tsouknidis, D. (2022). What matters for consumer sentiment in the euro area? World crude oil price or retail gasoline price? *Energy Economics, 105,* 105743. https://doi.org/10.1016/j.eneco.2021.105743
12. Cookson, J. A., & Niessner, M. (2020). Why don't we agree? Evidence from a social network of investors. *The Journal of Finance, 75*(1), 173-228. https://doi.org/10.1111/jofi.12852
13. Cordeiro, M., Sarmento, R. P., & Gama, J. (2016). Dynamic community detection in evolving networks using locality modularity optimization. *Social Network Analysis and Mining, 6*, 1-20. https://doi.org/10.1007/s13278-016-0325-1
14. Daoudi, A., Balouz, M., Bouakel, M., & Abada, A. (2025). Measuring the impact of digital technology on enhancing e-commerce in Algeria compared to a group of Arab countries (2010-2023). *Socio-Economic Relations in the Digital Society, 1*(55), 28-43. https://doi.org/10.55643/ser.1.55.2025.588
15. Daradkeh, M. K. (2022). A hybrid data analytics framework with sentiment convergence and multi-feature fusion for stock trend prediction. *Electronics, 11*(2). https://doi.org/10.3390/electronics11020250
16. Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys, 55*(13s), 1-38. https://doi.org/10.1145/3586075

DOI: 10.55643/fcaptp.5.64.2025.4823

17. Dzhamal, A., Kh. Sinh. (2025). EXAMINING FORENSIC ACCOUNTING'S ROLE IN SAFEGUARDING INDIAN BANKING INTEGRITY. *Financial and credit activity problems of theory and practice, 1*(60), 63-80. https://doi.org/10.55643/fcaptp.1.60.2025.4571

18. Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics, 9,* 1012-1031. https://doi.org/10.1162/tacl_a_00410

19. Fama, E. F. (1970). Efficient capital markets. *Journal of finance, 25*(2), 383-417. https://doi.org/10.7208/9780226426983-007

20. Gach, O., & Hao, J.-K. (2013). Improving the Louvain algorithm for community detection with modularity maximization. *International Conference on Artificial Evolution (Evolution Artificielle)*. https://doi.org/10.1007/978-3-319-11683-9_12

21. Gilbert, C. L., & Morgan, C. W. (2010). Food price volatility. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1554), 3023-3034. https://doi.org/10.1098/rstb.2010.0139

22. Gong, X., Guan, K., & Chen, Q. (2022). The role of textual analysis in oil futures price forecasting based on machine learning approach. *Journal of Futures Markets, 42*(10), 1987-2017. https://doi.org/10.1002/fut.22367

23. Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American economic review, 70*(3), 393-408. https://doi.org/https://www.jstor.org/stable/1805228

24. Guillaume, L. (2008). Fast unfolding of communities in large networks. Journal Statistical Mechanics: *Theory and Experiment, 10*, P1008. https://doi.org/10.1088/1742-5468/2008/10/P10008

25. Gurav, U. P., & Kotrappa, S. (2020). Sentiment aware stock price forecasting using an SA-RNN-LBL learning model. *Engineering, Technology & Applied Science Research, 10*(5), 6356-6361. https://doi.org/10.48084/etasr.3805

26. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

27. Iqbal, A., Amin, R., Iqbal, J., Alroobaea, R., Binmahfoudh, A., & Hussain, M. (2022). Sentiment analysis of consumer reviews using deep learning. *Sustainability, 14*(17), 10844. https://doi.org/10.3390/su141710844

28. Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of financial economics, 110*(3), 712-729. https://doi.org/10.1016/j.jfineco.2013.08.018

29. Kamara, A. F., Chen, E., & Pan, Z. (2022). An ensemble of a boosted hybrid of deep learning models and technical analysis for forecasting stock prices. *Information Sciences, 594*, 1-19. https://doi.org/10.1016/j.ins.2022.02.015

30. Kojaku, S., Radicchi, F., Ahn, Y.-Y., & Fortunato, S. (2024). Network community detection via neural embeddings. *Nature Communications, 15*(1), 9446. https://doi.org/10.1038/s41467-024-52355-w

31. Lemmon, M., & Portniaguina, E. (2006). Consumer confidence and asset prices: Some empirical evidence. *The Review of Financial Studies, 19*(4), 1499-1529. https://doi.org/10.1093/rfs/hhj038

32. Lin, Y., Chen, K., Zhang, X., Tan, B., & Lu, Q. (2022). Forecasting crude oil futures prices using BiLSTM-Attention-CNN model with Wavelet transform. *Applied Soft Computing, 130*, 109723. https://doi.org/10.1016/j.asoc.2022.109723

33. Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., & Gao, J. (2020). Adversarial training for large neural language models. *arXiv 2020*. https://doi.org/10.48550/arXiv.2004.08994

34. Liu, Y., & Matthies, B. (2022). Long-Run Risk: Is It There? *The Journal of Finance, 77*(3), 1587-1633. https://doi.org/10.1111/jofi.13126

35. Liu, Z., Zhou, B., Chu, D., Sun, Y., & Meng, L. (2024). Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion, 101*, 101973. https://doi.org/10.1016/j.inffus.2023.101973

36. O'hara, M. (2015). High frequency market microstructure. *Journal of financial economics, 116*(2), 257-270. https://doi.org/10.1016/j.jfineco.2015.01.003

37. Sanders, D. R., & Irwin, S. H. (2010). A speculative bubble in commodity futures prices? Cross-sectional evidence. *Agricultural Economics, 41*(1), 25-32. https://doi.org/10.1111/j.1574-0862.2009.00422.x

38. Sun, Y., Liu, Z., Sheng, Q. Z., Chu, D., Yu, J., & Sun, H. (2024). Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion, 110*, 102454. https://doi.org/10.1016/j.inffus.2024.102454

39. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139-1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

40. Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), 1437-1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

41. Wang, B., & Wang, J. (2020). Energy futures and spots prices forecasting by hybrid SW-GRU with EMD and error evaluation. *Energy Economics, 90*, 104827. https://doi.org/10.1016/j.eneco.2020.104827

42. Wang, Y., Pan, Z., Liu, L., & Wu, C. (2019). Oil price increases and the predictability of equity premium. *Journal of Banking & Finance, 102*, 43-58. https://doi.org/10.1016/j.jbankfin.2019.03.009

43. Wen, F., Gong, X., & Cai, S. (2016). Forecasting the volatility of crude oil futures using HAR-type models with structural breaks. *Energy Economics, 59*, 400-413. https://doi.org/10.1016/j.eneco.2016.07.014

44. Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR), 50*(2), 1-33. https://doi.org/10.1145/3057270

45. Yao, Jiaquan, Feng, Xu, Wang, Zanjun, Ji, Rongrong, & Zhang, Wei. (2021). Tone, mood and market impact: Based on the financial sentiment dictionary. *Journal of Management Science, 24*(5). https://doi.org/10.19920/j.cnki.jmsc.2021.05.002

46. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery, 8*(4), e1253. https://doi.org/10.1002/widm.1253

47. Zhang, W., Wu, J., Wang, S., & Zhang, Y. (2025). Examining dynamics: Unraveling the impact of oil price fluctuations on forecasting agricultural futures prices. *International Review of Financial Analysis, 97,* 103770. https://doi.org/10.1016/j.irfa.2024.103770

48. Zhang, Y., He, M., Wen, D., & Wang, Y. (2023). Forecasting crude oil price returns: Can nonlinearity help? *Energy, 262*, 125589. https://doi.org/10.1016/j.energy.2022.125589

49. Zhao, S., Hong, X., Yang, J., Zhao, Y., & Ding, G. (2023). Toward Label-Efficient Emotion and Sentiment Analysis This article introduces label-efficient emotion and sentiment analysis from the computational perspective, focusing on state-of-the-art methodologies, promising applications, and potential outlooks. Proc. IEEE. https://doi.org/10.1109/JPROC.2023.3309299

50. Zheng, Y., Li, X., & Nie, J.-Y. (2023). Store, share and transfer: Learning and updating sentiment knowledge for aspect-based sentiment analysis. *Information Sciences, 635*, 151-168. https://doi.org/10.1016/j.ins.2023.03.102

51. Zhou, Y., Fan, J., & Xue, L. (2024). How Much Can Machines Learn Finance from Chinese Text Data? *Management Science, 70*(12). https://doi.org/10.1287/mnsc.2022.01468

52. Zhukovska, A. (2025). Genesis of scientific concepts of inclusive economic development. *Socio-Economic Relations in the Digital Society, 1*(55), 5-18. https://doi.org/10.55643/ser.1.55.2025.594

53. Zhu, X., Zhu, L., Guo, J., Liang, S., & Dietze, S. (2021). GL-GCN: Global and local dependency guided graph convolutional networks for aspect-based sentiment classification. *Expert Systems with Applications, 186,* 115712. https://doi.org/10.1016/j.eswa.2021.115712

54. Zhu, Z., & Mao, K. (2023). Knowledge-based BERT word embedding fine-tuning for emotion recognition. *Neurocomputing, 552*, 126488. https://doi.org/10.1016/j.neucom.2023.126488

Гао Ц., Юй Х.

ВИВЧЕННЯ РОЛІ ТЕКСТОВОГО АНАЛІЗУ В ПРОГНОЗУВАННІ ЦІН НА Ф'ЮЧЕРСИ НА СОЄВУ ОЛІЮ

Автори статті пропонують інноваційну структуру, яка поєднує текстову аналітику з кількох джерел із машинним навчанням для розв'язання серйозної проблеми інформаційної асиметрії на ринку деривативів. Щоб розглянути недоліки традиційних структурованих моделей даних у фіксації спекулятивної торгової поведінки, автори кількісно оцінюють механізм передачі настроїв інвесторів на ф'ючерсний ринок за допомогою синергетичного моделювання з обробкою природної мови та складним мережевим аналізом. По-перше, побудовано метод розширення словника настроїв для предметної області ф'ючерсів на основі моделі BERT, який розв'язує проблему недостатнього охоплення ф'ючерсних термінів у словниках загального призначення. По-друге, автори будують динамічну мережу стосунків із користувачами за допомогою алгоритму Лувена, інтегруючи три модальні особливості взаємодії: схожість контенту, синхронізацію часу та актуальність атрибутів, – щоб виявити структурні моделі еволюції високочастотних торгових груп і спільнот технічного аналізу в барі ф'ючерсів на соєву олію. Нарешті вони розробляють механізм оптимізації узгодженості заголовків і текстів на основі Generative Adversarial Network (GAN), щоб динамічно виявляти тексти, що конфліктують за настроями, за допомогою архітектури Transformer-CNN-MLP і заповнювати відсутні значення шляхом комбінації зі стратегією нейтрального заповнювача. Емпіричні дані показують, що модель LSTM, яка поєднує новий лексикон настроїв і метрики впливу спільноти зі змагальними вагами узгодженості, має найкращі прогнозні показники порівняно з іншими еталонними моделями. Цей досвід забезпечує інтерпретований технічний шлях для фінансового прогнозування на основі соціальних медіа шляхом синергії текстових неявних функцій із моделями групової взаємодії, а його модульний дизайн може бути поширений на царину управління товарними ризиками та RegTech.

Ключові слова: фінансові деривативи, аналіз неструктурованих фінансових даних, словник фінансових настроїв, ціноутворення ф'ючерсів на сільськогосподарську продукцію, моніторинг впливу на громаду, стабільність ринку, управління ризиками

JEL Класифікація: G11, Q11, C53

DOI: 10.55643/fcaptp.5.64.2025.4823