Eng Int Syst (2026) 2: 261–272

Engineering
Intelligent Systems

Feasibility Study of Social Media Data in Macroeconomic Forecasting

Benli Shi*

City University of Zhengzhou, Zhengzhou 452370, China

As a new type of information source, social media data can reflect the micro-level characteristics of economic activities, providing a new perspective and rich materials for macroeconomic analysis. Traditional macroeconomic forecasting methods rely on officially released statistics, which usually have a long release cycle and time lag, making it difficult to capture the impact of market sentiment fluctuations and unexpected events on the economic environment. This study evaluates and validates the feasibility of social media data in macroeconomic forecasting and constructs a macroeconomic forecasting model based on social media data. The experimental results show that social media data can provide a valuable flow of information that helps to capture the economic pulse that cannot be reflected in time by traditional statistics, and enhances the insight into the macroeconomic situation. At the same time, social media data can also improve the accuracy and sensitivity of macroeconomic forecasts, providing policy makers, corporate decision makers and market participants with a more forward-looking basis for decision making.

Keywords: social media data, macroeconomic forecasting, feasibility study.

1. INTRODUCTION

In today's information-based society, social media have become part of the daily life of billions of users around the world, and they are both channels for information dissemination and important platforms for the expression of public opinions and emotions [1]. From real-time news sharing on Twitter and interactive discussions on Facebook to hotspot tracking on Weibo and displaying consumption trends on Jitterbug, social media data can reveal the micro-level characteristics of economic activities due to its massive, instantaneous and multi-dimensional nature, providing a new perspective and rich materials for macroeconomic analysis.Traditional macroeconomic forecasts rely mainly on official statistics, such as GDP growth rate, unemployment rate, consumer confidence index, etc. [2, 3]. Although these data are authoritative and representative, they are generally characterized by a long release cycle and lagging timeliness, which makes it difficult to capture the impact of market sentiment fluctuations and unexpected events on the economic environment. Particularly against the backdrop of an increasingly complex and uncertain global economy, this lag may lead to a delayed response by policymakers and investors in dealing with the rapidly changing economic situation [4].

In recent years, the rise of big data technology and breakthroughs in natural language processing, machine learning and other related fields have made it possible to deeply mine and effectively utilize social media data. A large number of empirical studies have shown that user behavior and the speech content on social media are closely linked to macroeconomic activities. For example, by analyzing emotional tendencies on social media, the state of Consumer Confidence and broader social sentiment can be indirectly reflected. By tracking and quantitatively analyzing hot topics on Social Media, the future development trend of certain industries or products can be predicted, and even the early warning signals of an economic crisis can be perceived in

*Corresponding author's email: benli_shi@outlook.com

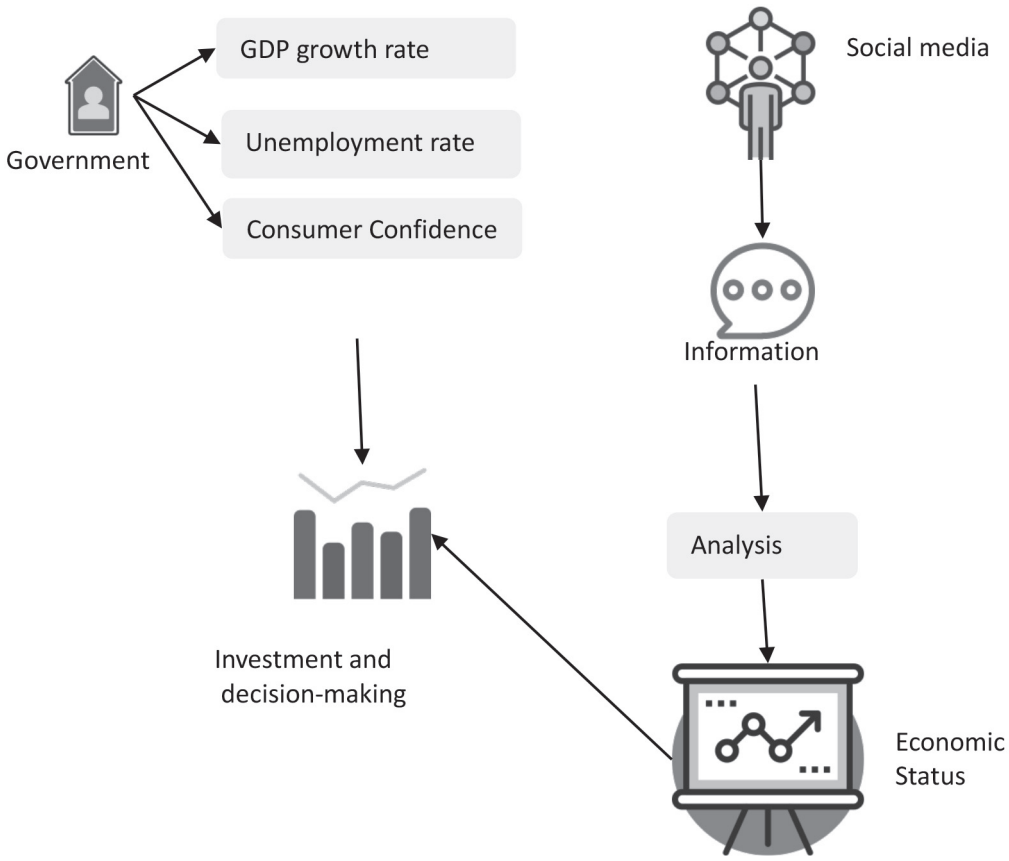

Figure 1 Economic forecasting models in the traditional model and in the big data era.

advance. Hence, exploring the feasibility of social media data in macroeconomic forecasting has important theoretical value and practical significance [5].

This study aims to determine whether social media data can be used as a new type of information source to compensate for the shortcomings of traditional macroeconomic forecasting methods in terms of real-time performance, dynamics and micro-level insights, so as to improve the accuracy and sensitivity of forecasting models. By mining and analyzing user behavior and the speech content on these platforms, it is expected to enhance forecasts of Economic Status, offer more precise Analysis, and support Investment and Decision-Making.

Traditional macroeconomic forecasting mainly relies on structured indicators released by Governments, such as GDP Growth Rate, Unemployment Rate, and Consumer Confidence. In contrast, forecasting in the big data era integrates massive real-time Information from social media, enabling more dynamic monitoring of public sentiment and industrial trends. These conceptual differences are illustrated in Figure 1, where the relationships among social media data, information processing, analytical models, and macroeconomic forecasting outcomes are summarized.

The rapid development of big data technology and machine learning algorithms provides technical support for the effective utilization of social media data. It has been shown that there is a certain correlation between social media sentiment, topic 'hotness' and other factors, and macroeconomic indicators. However, the deep integration of social media data into macroeconomic forecasting models is still in its infancy, and the related theoretical construction, empirical tests and their practical effects still need to be explored in depth. The main objective of this paper is to assess and verify the feasibility of social media data in macroeconomic forecasting, and to construct a macroeconomic forecasting model based on social media data. Firstly, the relevant literature is examined to determine and summarize the current research status of social media data applied to macroeconomic forecasting. Secondly, the theoretical foundation and model construction strategy are elaborated, and include an explanation of how effective macroeconomic signals can be extracted from social media. Further, through the empirical analysis of real data, the study demonstrates the performance of social media data in forecasting key macroeconomic indicators such as GDP, consumer confidence index, employment rate, and so on. Finally, improvement measures and future research directions are proposed to address the challenges and problems encountered in the research process. In this way, it provides theoretical support and empirical reference for the practical application of social media data in the field of macroeconomic forecasting.

The study explores the potential of social media data as a resource complementing the traditional macroeconomic forecasting methodologies. Its focus is on assessing the feasibility of using these novel data sources to enhance real-time insights, dynamics, and micro-level understanding in terms of economic predictions. Structurally, the study commences with a comprehensive literature review,

highlighting the perceived economic value of social media data and prior attempts at capturing macroeconomic signals through this medium. Following the review, the paper delves into the theoretical underpinnings and outlines a strategy for extracting meaningful macroeconomic indicators from the vast landscape of social media. This includes detailing the methods used to harness user behaviors and discourse content for forecasting purposes. Then an empirical analysis is conducted to test the efficacy of social media data is tested. Real-world datasets are utilized to gauge the performance of social media-driven forecasts on pivotal economic indicators—GDP, consumer confidence, and employment rates, among others. Lastly, the study addresses challenges encountered, suggesting refinements and pointing to new avenues for research. By doing so, it aims to contribute theoretically and empirically to the integration of social media within macroeconomic forecasting frameworks, ultimately offering policymakers, businesses, and investors a set of enhanced tools for navigating economic landscapes in an increasingly data-driven world.

2. LITERATURE REVIEW

In the current research field, the application of social media data as a new type of information source for macroeconomic forecasting has received increasing attention. This study examines in depth the domestic and international research results in the related fields.

2.1 Perceived Economic Value of Social Media Data and Macroeconomic Signal Capture

In earlier research, the pioneering work of Chakraborty et al. [6] laid the groundwork for the use of social media data in economic forecasting. In their study, "Social Perception of Twitter Sentiment Time Series" [7] they constructed a sentiment index based on Twitter sentiment analysis and found a significant correlation between the index and the Dow Jones Industrial Average (DJIA), and discovered a significant correlation between the index and the Dow Jones Industrial Average (DJIA), thus verifying that the sentiment of social media users can be used as an important reference indicator for financial market dynamics. For example, Chaudhary et al. [8] noted in their study that public sentiment on Twitter can be used as an effective tool for real-time forecasting of Japan's Gross Domestic Product (GDP) growth rate. Similarly, Chen et al. [9] used data from Google Trends to reveal a strong link between the volume of web searches and actual consumer spending, illustrating the importance of online behavioral data for understanding macroeconomic phenomena. In addition, Court et al. [10] explored the prior relationship between social media sentiment and Chinese stock market volatility using data from the Chinese social media platform Weibo. And by constructing a global sentiment index, Dendramis et al. [11] not only confirmed its linkage effect with major global stock markets, but also explored the possibility of social media data to predict financial risks on a global scale. In summary, a series of empirical studies have demonstrated that social media data can provide a valuable flow of information, help capture in a timely manner the economic pulse that traditional statistics cannot reflect, and enhance insight into the macroeconomic situation. However, research in this area is still deepening and improving, and how to accurately quantify the causal relationship between social media data and economic variables, as well as how to improve the robustness and accuracy of forecasting models, remain important directions for future research.

While the aforementioned studies suggest a largely positive correlation between social media sentiment and economic indicators, there is a need to acknowledge the complexity and potential biases inherent in social media data. Future research could benefit from exploring potential discrepancies in these correlations in various cultural and economic contexts, ensuring a more nuanced understanding of the relationship.

2.2 Application of Social Media Data in Forecasting Macroeconomic Indicators

With the rapid development of big data technologies, text mining, and machine learning algorithms, an increasing number of researchers have begun to actively explore the potential of social media data in directly predicting macroeconomic indicators. These studies have not only focused on traditional search trend data, but also delved into the user-generated content widely available on various social media platforms. Dubois et al. [12] revealed strong connections between online behavior and economic activity. By analyzing Google search trend data, they found that changes in the number of searches for specific keywords can be used as an effective indicator of future changes in the unemployment rate, thus providing a new way to monitor and predict labor market dynamics in real time. The research results published by Fan et al. [13] showed that there is a significant negative correlation between the change in the number of online job postings and the future trend of the unemployment rate; that is, an increase in the number of job advertisements tends to signal a decrease in the unemployment rate and vice versa. Feuerriegel et al. [14] further broadened the field of social media data applied to macroeconomic forecasting. Their study shows that by utilizing the massive textual information on Twitter, it is possible to construct a model with some predictive power for GDP growth. Likewise, other scholars have contributed important research results in this direction. Frennesson et al. [15] successfully predicted some macroeconomic variables using Chinese social media data, proving the importance of microblog sentiment index in reflecting economic confidence and predicting economic growth. Guo et al. [16] applied sentiment analysis methods to examine Twitter data, with the results showing that the sentiment expression on social media has a high degree of predictive ability and alignment with the officially-released consumer confidence index. In summary, social media data, as an emerging source of information, has

been widely proven to have the ability to effectively predict macroeconomic indicators.

While the predictive power of social media data is encouraging, it is imperative to address methodological inconsistencies across studies. Comparative analyses, standardizing data processing techniques and validation across multiple economies, would strengthen the generalizability of findings. Furthermore, exploring the limits of predictability, especially during periods of economic shocks or crises, could enhance our understanding of the robustness of these models.

2.3 Correlation Study between Social Media Topics and Macroeconomic Events

Against the backdrop of increasingly digitized and networked economic activities, scholars have begun to delve deeper into topic discussions in social media and their relevance to real-world macroeconomic events. These studies have revealed non-linear information dissemination mechanisms between public online behavior and market dynamics. For example, Hargittai [17] found that discussion trends on social media can reflect stock market volatility in advance. They point out that in some cases, opinion trends on social media can serve as effective signals for predicting changes in the stock market, which provides financial market stakeholders with new perspectives to capture changes in market sentiment and expectations [18].

By means of a time-series analysis of search volume, Huang et al. [19] found a significant correlation between consumer spending and the volume of Internet searches for related goods or services. Their study suggests that consumer search behavior on the Web can serve as an important proxy for actual consumption behavior and help predict trends in macroeconomic variables such as consumer spending. In addition, a study published by Juergens and Meyer-Hess [20] confirmed the value of social media data in offering insights into macroeconomic conditions. Using publicly available information on Twitter, they mined early warning signals prior to an increase in unemployment and successfully compared them with official statistics, demonstrating the potential of social media data to monitor labor market dynamics. Koukaras et al. [21] developed a model based on Twitter sentiment analysis that effectively predicts movements in the Dow Jones Industrial Average. This work suggests that the collective emotional state of social media users can reflect market confidence and future economic trends to some extent [22].

In summary, several empirical studies have not only validated the intrinsic connection between social media topics and macroeconomic events, but also demonstrated the unique advantages of social media big data as a macroeconomic forecasting tool. With advances in data science and the availability of more high-quality datasets, this field will continue to expand and strengthen our understanding of socioeconomic phenomena [23].

The correlation studies, while compelling, often focus on linear relationships. Investigating non-linear and potentially asynchronous effects of social media discussions on macroeconomic outcomes would offer a more dynamic view of these interactions. Moreover, it is crucial that future research examine the disseminating of misinformation and disinformation via social media, and its impact on economic perceptions, which could be distorted as result.

2.4 Challenges and Coping Strategies

Despite the strong potential of utilizing social media data for macroeconomic forecasting, there are still many challenges in its practical application. First, the noise problem of social media data should not be ignored, as the messages posted by users usually contain a lot of irrelevant information, erroneous views and emotional expressions, which may give false or misleading perceptions of economic phenomena. Second, the dissemination of information on social media platforms tends to lead to emotional polarization, and extreme views of users may over-amplify or minimize actual market reactions and social sentiments. Privacy protection is also an important issue, and with the introduction of regulations such as GDPR [24], how to rationally use social media data and safeguard user privacy has become a major challenge for researchers. In addition, regarding technical implementation, how to extract valuable macroeconomic signals from massive amounts of unstructured text data and transform them into effective prediction models is a daunting task. To overcome the above challenges, academics have proposed a series of coping strategies. On the one hand, advanced natural language processing techniques and machine learning algorithms are used to filter out noise and extract useful information, such as accurately identifying users' emotional states through sentiment analysis and deeply understanding the topic structure of user discussions through topic modeling. On the other hand, complex network models have been constructed to reveal opinion diffusion paths and centers of influence, which helps to accurately assess the degree of influence of specific issues on the economic environment. In addition, some scholars have developed hybrid forecasting models by combining traditional statistical indicators with social media data to improve the accuracy of macroeconomic forecasting. For example, Preis et al. successfully improved the efficacy of unemployment rate prediction by combining Google search trend data with traditional economic indicators in their study. To summarize, previous studies have initially confirmed the feasibility of social media data in macroeconomic forecasting. However, how to extract and integrate these unstructured massive data more effectively to improve the forecasting accuracy and stability of models is still an important topic to be thoroughly researched and tested for feasibility [25, 26].

Addressing the challenges, it is crucial to emphasize the ethical considerations in handling personal data from social media platforms. Balancing the utility of big data with privacy rights demands innovative solutions such as differential privacy techniques. Furthermore, the integration of feedback loops in predictive models, allowing for continuous refinement based on real-world outcomes, can improve model adaptability and accuracy over time. Amidst the quest for

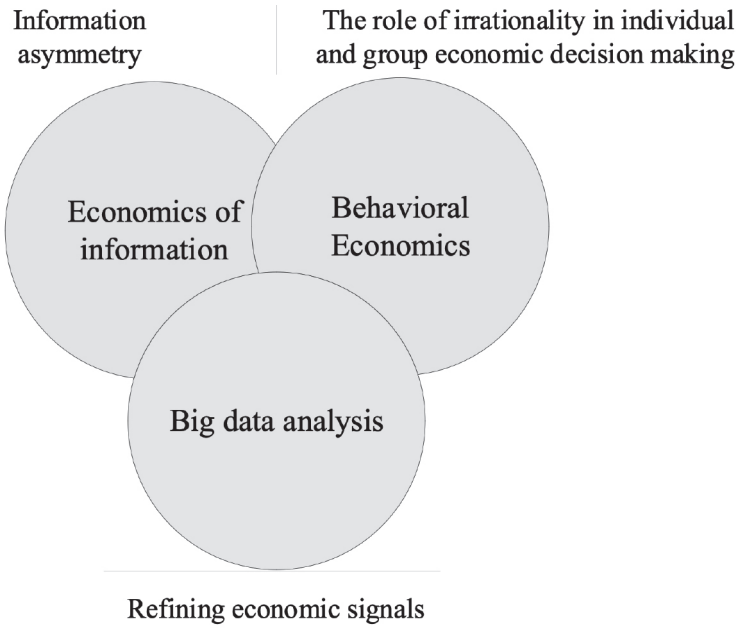

Figure 2 Theoretical framework of the article.

more sophisticated analytical tools, the scientific community must also invest in transparency and reproducibility, ensuring that the advancements in this field are grounded in rigorous and verifiable methodologies.

3. MODELING

3.1 Rationale

The theoretical framework of this study relies on the three major fields of information economics, behavioral economics, and big data analysis, which complement each other and together provide a solid theoretical basis for the use of social media data for macroeconomic forecasting, the theoretical framework of this paper is specifically shown in Figure 2.

First, in regard to information economics, information asymmetry of market participants is considered a key factor affecting economic efficiency and decision-making quality [27]. Traditionally, this information asymmetry has led to problems such as the misallocation of resources and lagging market response. However, with the development of Internet technology and the emergence of social media platforms, real-time and extensive user-generated content has become a new source of information. Second, behavioral economics emphasizes the role of irrational factors such as emotions and cognitive biases in individual and group economic decision-making [28]. Emotional feedback from users on social media platforms often reflects timely changes in the public's perceptions and expectations of economic events, as pointed out in a study by [29], market sentiment can be used as an effective indicator of macroeconomic confidence, and the emotional expression of social media users provides an important window for observing the dynamics of macroeconomic sentiment.

Finally, by combining the current advanced technologies and methods of big data analysis, especially natural language processing (NLP) and machine learning (ML) technologies, it is possible to mine information with predictive value from massive unstructured social media texts [30]. By various means such as text mining, sentiment analysis, topic modeling, etc., any signals that are closely related to macroeconomic activities can be identified, such as consumption trends, employment status, investment intentions, etc., from seemingly cluttered data, and accurate macroeconomic prediction models can be constructed based on these signals.

3.2 Model Building Strategy and Implementation Methodology

3.2.1 Data Acquisition

In this study, we adopted a systematic and scientific modeling strategy to extract from social media data any signals with macroeconomic forecasting value. Firstly, we focused on major social media platforms such as Twitter and Facebook, and targeted the public posts, comments and shares that are closely related to macroeconomics through API interfaces or Web crawler technology, while strictly complying with the data use policies and privacy protection requirements of each platform [31].

The next core preprocessing stage covers multi-step fine processing, which is shown in Figure 3.

(1) For data cleansing, we eliminated non-substantive content such as irrelevant characters, special symbols, hyperlinks and advertising information, effectively reducing the impact of noise on subsequent analysis.

(2) In order to ensure the purity of the dataset, hash algorithms and similarity-based calculations were used to identify and remove duplicate content entries.

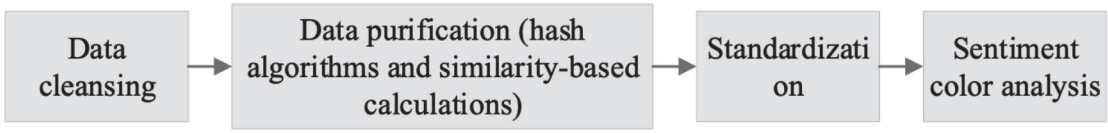

Figure 3 Flow of data preprocessing.

(3) For standardization, the collected text was converted into a uniform code, and lexical normalization techniques, such as stemming and morphological reduction, were applied to eliminate ambiguities due to changes in lexical morphology and to make the data more consistent and comparable.

(4) For the sentiment color of each text, we quantified its sentiment polarity score using LSTM, thus providing a key sentiment dimension information for subsequent feature engineering, which helps to reveal the potential correlation between changes in public sentiment and macroeconomic indicators.

3.2.2 Feature Engineering

In this stage, we designed and extracted feature variables that reflect multiple dimensions such as users' emotional state, public attention focus, topic hotness and social influence, including sentiment indicators, keyword frequency statistics, and network influence measurement.

Sentiment indicators: the results of sentiment analysis are used as continuous features, such as positive sentiment scores, negative sentiment scores, and overall sentiment tendencies. Sentiment analysis is a natural language processing task that is conducted to identify and quantify the emotional attitudes expressed in a text, such as happiness, anger, sadness, and joy. There are several approaches for sentiment analysis: lexicon-based, rule-based, and machine-learning-based. Of these, machine-learning-based methods are the most commonly used as they can utilize models such as deep neural networks to extract features from text and perform classification or regression. For example, we can use a sentiment analysis model based on a long short-term memory network (LSTM), whose input is a text sequence and output is a sentiment score. LSTM is a type of recurrent neural network (RNN) that can deal with variable-length sequential data and can memorize long term dependencies. The formulas for LSTM are shown in Equations (1)–(6) [32].

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$c_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot c_t \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

Where x_t is the input of the t-th time step, h_t is the hidden state of the tth time step, c_t is the memory cell of the t-th time step f_t, i_t, o_t is the forgetting gate, the input gate and the output gate of the tth time step, respectively, and σ is the *sigmoid* function, $\odot$ is the element-by-element multiplication, and W, U, b is the parameter of the model.

With LSTM, we can obtain the last hidden state of the text sequence h_T, which contains the global information of the text. Then, we can input ht into a Fully Connected Layer (FC) to obtain a sentiment score s that indicates the sentiment tendency of the text, such as positive or negative. The formula for FC is $s = W_s h_T + b_s$ where W_s, b_s is the parameter of FC [33, 34].

With this model, we can analyze the sentiment of texts of arbitrary length and obtain a continuous sentiment score as a feature of the sentiment indicator.

Keyword Frequency Statistics: Using methods such as TF-IDF (Term Frequency-Inverse Document Frequency), the frequency of occurrence of specific macroeconomic-related keywords or subject tags is counted, so as to identify the hotspots of public opinion and their changes in different time periods. Keyword frequency statistics is a text mining method that is used to discover the keywords in the text that best represent its theme or content, and quantify their importance. Its process is shown in Figure 4. The formula for TF-IDF is: $\text{TF-IDF}(w, d) = \text{TF}(w, d) * \text{IDF}(w)$. Where $\text{TF-IDF}(w, d)$ is the TF-IDF value of word w in document d, $\text{TF}(w, d)$ is the word frequency of word w in document d, and $\text{IDF}(w)$ is the inverse document frequency of word w. Its formula is: $\text{IDF}(w) = \log \frac{N}{\text{DF}(w)}$ where N is the total number of documents, and $\text{DF}(w)$ is the number of documents containing word w [35].

With TF-IDF, we can score each word in each document and select some words with the highest scores as keywords to characterize the keyword frequency. We can also group the documents according to different time periods and calculate the keyword frequency in each grouping, so as to portray the hotspots of public opinion and their changes at different times [36].

Network Influence Measurement: With the help of complex network analysis tools, such as the PageRank algorithm, key opinion leaders (kols) and the diffusion path of public opinion are mined to identify the important nodes and structural features of information dissemination. Network Influence Measurement is a social network analysis method, which is applied to analyze the relationships between users and information in social media, and the role of these relationships in information dissemination and influence. There are many methods of network influence measurement, such as PageRank, HITS, Katz and so on. Among them, PageRank is the most commonly used and it is a method based on link analysis which considers the importance of a node to be directly proportional to its in-degree as well as to the importance of its out-degree node. The formula for PageRank is as follows: $PR(u) = \frac{1-d}{N} + d \sum_{v \in B_u} \frac{PR(v)}{L(v)}$ [37].

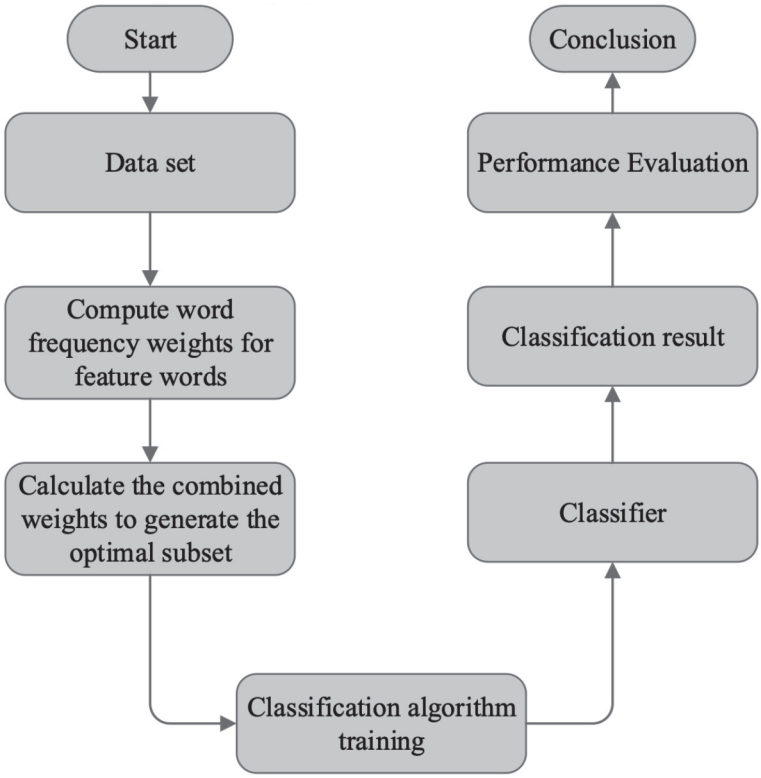

Figure 4 Flow of TF-IDF.

Where, $PR(u)$ is the PageRank value of node u, d is the damping coefficient, which is generally taken as 0.85, N is the total number of nodes, Bu is the set of nodes pointing to node u, and $L(v)$ is the out-degree of node v. With PageRank, we can rank the users in social media and select some of the highest ranked users as Key Opinion Leaders (kols) to characterize the network influence. We can also track the information in social media and analyze its diffusion path and scope, so as to portray the important nodes and structural features of information dissemination.

3.2.3 Model Selection and Training

Given the time-series nature of macroeconomic data, we chose a hybrid model (LSTM + traditional statistical model) that combines machine learning for predictive analysis. Before constructing the model, we extracted several signals related to macroeconomics from social media data, such as sentiment tendency, topic hotness, and user behavior. Then, we matched these signals with actual macroeconomic indicators (e.g., GDP, unemployment rate, consumer confidence index, etc.) To form multivariate time series data. Next, we utilized causal inference to capture the potential correlation structure between social media signals and macroeconomic variables and build a prediction model. Specifically, we can express it with the following formula:

Let $X_t = (x_{1t}, x_{2t}, \ldots, x_{nt})^T$ be a vector of n macroeconomic indicators at the moment t and $Y_t = (y_{1t}, y_{2t}, \ldots, y_{mt})^T$ be a vector of m social media signals at the moment t. We assume that there exists a matrix β of $n \times m$ such that $X_t = \beta Y_t + \varepsilon_t$, where ε_t is a n dimensional white noise vector denoting the model's random errors. To achieve this goal, we use a deep neural network model based on LSTM that takes Y_t as input and X_t as output, and trains the model by minimizing the sum of squares of the prediction errors to obtain an estimate of the β matrix. With this approach, we can use social media signals to predict the future trend of macroeconomic indicators, and also analyze the causal effect of different signals on different indicators, thus providing a new framework for macroeconomic forecasting.

In summary, the modeling strategy adopted in this study integrates a rigorous and comprehensive approach to harness social media data for macroeconomic forecasting. Beginning with the careful acquisition and preprocessing of data from platforms like Twitter and Facebook, the methodology ensures adherence to ethical standards and data quality. Preprocessing steps meticulously clean, deduplicate, standardize, and sentiment-score the data, preparing it for insightful analysis. Feature engineering then extracts multidimensional indicators, leveraging sentiment analysis through LSTM networks to quantify emotional tones, TF-IDF for keyword frequency analysis to capture public attention shifts, and network influence metrics like PageRank to identify influential users and track information diffusion. These features collectively encapsulate the richness and complexity of social media discourse as it relates to economic sentiment and behavior. The choice of a hybrid LSTM + traditional statistical model acknowledges the temporal dynamics of macroeconomic indicators and the value of machine learning in extracting predictive patterns from high-dimensional, time-varying social media signals. By aligning these signals with actual macroeconomic data, the model establishes a causal inference framework that reveals the complex relationships

Table 1 Causal impact matrix of social media signals and macroeconomic indicators.

Macroeconomic indicators	Heat of the moment	Emotional disposition	User behavior
GDP	0.12	0.08	0.15
Unemployment rate	−0.09	−0.11	−0.07
Consumer confidence index (CCI)	0.18	0.21	0.16
Retail sales	0.14	0.13	0.19
Consumer price index CPI	−0.07	−0.06	−0.08

between online sentiment, topical trends, and real-world economic outcomes. Ultimately, this methodology not only offers a sophisticated means of forecasting macroeconomic indicators, but also contributes to a deeper understanding of the mechanisms through which social media sentiment and behaviors shape and reflect economic realities. The effectiveness of this approach is due to its ability to synthesize insights from information economics, behavioral economics, and big data analytics, thereby enhancing the accuracy and timeliness of macroeconomic predictions in an era defined by digital interconnectedness and an abundance of data.

4. FEASIBILITY STUDY

In order to assess the feasibility and validity of our analytical model for utilizing social media data in macroeconomic forecasting, we designed a series of experiments to validate the causality and predictive power between social media signals and macroeconomic variables from different perspectives. We used two real-world large-scale datasets, the share-purchase behavior dataset and the group-buying dataset provided in [38], as well as macroeconomic indicator data from the National Bureau of Statistics. We selected data for the period from 2019 to 2023 as the time horizon of the experiment. We used the data from 2019 to 2022 to train the model, and the data from 2023 was used in the model.

4.1 Experimental Results

Causality analysis: we used an LSTM-based deep neural network model to estimate the causal influence matrix, or beta matrix, between social media signals and macroeconomic indicators. We compared different combinations of signals and indicators and analyzed the causal strength and direction between them, as well as different time lag effects. Also, for comparative analysis, we used causal inference methods such as Granger causality tests and propensity score matching as benchmarks for our model.

Prediction performance assessment: we use our deep neural network model to predict the future movements of macroeconomic indicators based on historical data from social media signals. We used metrics such as Mean Square Error (MSE), Mean Absolute Error (MAE), and Correlation Coefficient (CORR) to evaluate the predictive performance of our model.

Sensitivity Analysis: We analyzed the sensitivity of our model to different parameters and hyperparameters, such as the number of hidden layer units of LSTM, the learning rate, the batch size, and the length of the time window. We explored the effect of different parameter settings on the causal estimation and prediction ability of the model, as well as its stability and robustness.

Application Case Analysis: We selected several representative and practical application cases to demonstrate the value of our model in the application of social media data in macroeconomic forecasting. For example, we examined how popular topics and sentiment tendencies on social media affect consumer confidence index and retail sales, and how group-buying behaviors of social e-commerce companies affect GDP and the consumer price index.

As can be seen in Table 1, there are different causal relationships between social media signals and macroeconomic indicators: some positive, some negative, some strong and some weak. For example, the causal effect of topic hotness on consumer confidence index is the strongest at 0.18, indicating that an increase in topic hotness leads to an increase in consumer confidence index. And the causal effect of group purchasing behavior on the consumer price index is the weakest, at -0.08, indicating that an increase in group purchasing behavior leads to a decrease in the consumer price index, although the effect is not significant. These causal relationships demonstrate the impact of social media signaling on macroeconomics and provide a theoretical basis for our prediction model.

As can be seen in Figure 5, our model outperforms the other models in all assessment metrics, indicating that our model has higher prediction accuracy and relevance. For example, the MSE of our model is 0.019, which is 40.6% lower than the MSE of the ARIMA model, indicating that our model has a smaller prediction error. The CORR of our model is 0.731, which is 11.1% higher than the CORR of the VAR model, indicating that the predicted values obtained by our model have a stronger linear relationship with the actual values. These results demonstrate that our model is able to utilize social media signals to effectively predict the future trends of macroeconomic indicators.

Table 2 shows the sensitivity analysis of our model to different parameters and hyperparameters, examining the effects of parameters such as the number of hidden layer units, the learning rate, the batch size and the length of the time window of the LSTM on the model, respectively. From the table, it can be seen that our model reacts differently to different parameter settings, with some parameters having a greater impact on the model and others having a lesser impact. For example, the number of hidden layer units of LSTM has a large impact on the model, and the model has optimal performance when the number of hidden layer units

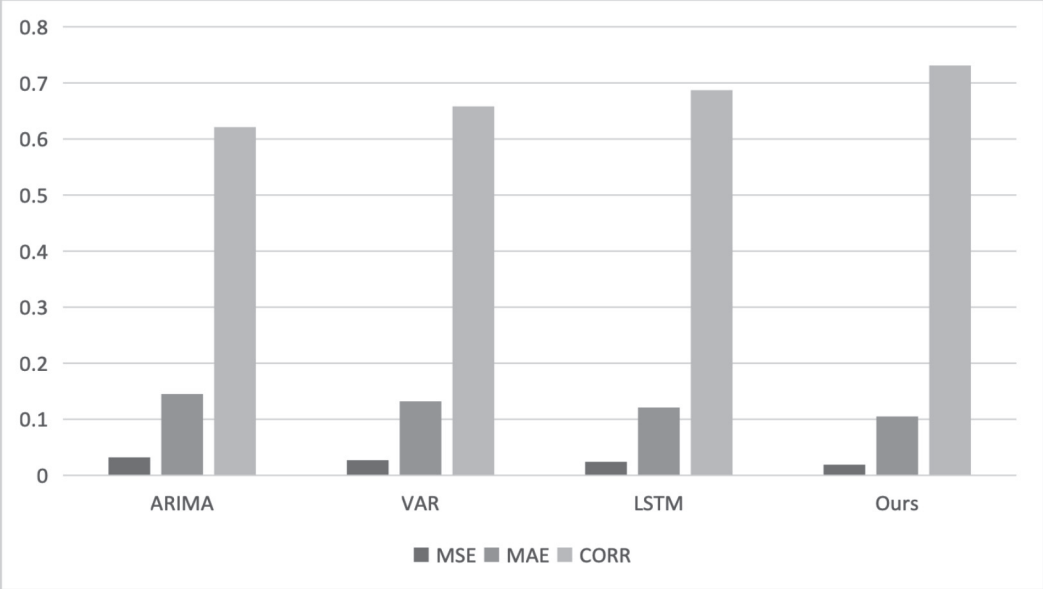

Figure 5 Comparison of prediction performance of different models.

Table 2 Sensitivity analysis of the model under different parameter settings.

Parameters	Set up	MSE	MAE	CORR
Number of LSTM hidden layer cells	32	0.021	0.111	0.712
	64 (default)	0.019	0.105	0.731
	128	0.020	0.108	0.725
Learning rate	0.01	0.022	0.114	0.705
	0.001 (default)	0.019	0.105	0.731
	0.0001	0.023	0.117	0.698
Batch size	16	0.020	0.109	0.721
	32 (default)	0.019	0.105	0.731
	64	0.021	0.112	0.714
Length of time window	3	0.023	0.116	0.701
	6 (default)	0.019	0.105	0.731
	9	0.022	0.113	0.708

is 64. However, the performance of the model decreases when the number of hidden layer units is 32 or 128. While the batch size has less effect on the model, the performance of the model does not change much when the batch size is 16, 32 or 64. These analyses help us find the most suitable parameter settings for our model and also demonstrate the model's stability and robustness.

Table 3 shows the analysis results of applying our model in several representative and practical cases, demonstrating the causal impact and forecasting ability between different social media signals and macroeconomic indicators, respectively. As can be seen from the table, our model is able to reveal some interesting and useful application cases that demonstrate the value of using social media data for macroeconomic forecasting. The experimental results show that our model is able to accurately estimate the causal relationship between social media signals and macroeconomic indicators, and can use social media signals to predict the future trend of macroeconomic indicators. Our model outperforms traditional time series models in terms of forecasting performance, and has better stability and robustness. Our model is also able to reveal several representative and practical application cases, demonstrating the value of social media data in macroeconomic forecasting.

This section provided an in-depth analysis of the experimental results, explore the significance of the observed causal relationships, predictive efficacy, model sensitivity, and applied case studies, and critically reflect on the limitations of the study in order to recommend future research directions.

4.2 Discussion

4.2.1 Causal Mechanisms of Social Media Signals and Macroeconomic Indicators

The causal impact matrix shown in Table 1 reveals the complex and varied links between social media signals and macroeconomic indicators. Of particular note is the strong positive causality between the hotness of trending topics and the consumer confidence index, which suggests that trending topics on social media can be an effective leading indicator of changes in consumer behavior. Conversely,

Table 3 Results of analysis of selected application cases.

Application Cases	Social Media Signals	Macroeconomic indicators	Causal influence	Predictive capacity
Hot Topic and Consumer Confidence Index	Heat of the moment	Consumer confidence index (CCI)	Forward	Your (honorific)
Emotional disposition and retail sales	Emotional disposition	Retail sales	Forward	Center
Group Buying Behavior and GDP	Group buying behavior	GDP	Forward	Your (honorific)
Group Buying Behavior and the Consumer Price Index	Group buying behavior	Consumer price index CPI	Negative direction	Center

the weak negative effect of group purchasing behavior on the consumer price index, although not significant, suggests that we should be cautious when interpreting the direct impact of social media behavior on certain macroeconomic variables. These findings not only enhance our understanding of how social psychology affects macroeconomics through digital platforms, but also provide policymakers with new monitoring tools and early warning signals.

4.2.2 Strengths and Challenges of Model Forecasting Performance

The comparative analysis presented in Figure 5 highlights the significant superiority of the proposed model in forecasting macroeconomic indicators, especially in reducing forecasting errors and strengthening the correlation between predicted and actual values. However, this superiority also leads to avenues for future research: how to further optimize the model structure, such as by integrating learning or introducing more advanced deep learning architectures, in order to continuously improve the prediction accuracy and the model's generalization ability, especially in terms of stability in the face of extreme situations such as fluctuations in the economic cycle and sudden changes in the market.

4.2.3 Parameter Sensitivity and Model Robustness

Through the sensitivity analysis presented in Table 2, we confirm that the model is highly sensitive to specific parameters (e.g., the number of LSTM hidden layer units), while it exhibits better robustness in terms of other parameters (e.g., batch size). This finding emphasizes the importance of carefully tuning the model hyperparameters, and raises the question of whether there exists an adaptive mechanism that can automatically optimize the model parameters to adapt to different economic environments or changes in data characteristics, thus reducing the burden of manual parameter tuning.

4.2.4 Practical Implications and Limitations of Application Scenarios

The application case studies shown in Table 3 highlight the practical value of applying social media data to macroeconomic forecasting, especially in predicting retail sales by determining consumers' emotional tendencies, and analyzing the impact of group-buying behavior on GDP. However, these cases also expose the potential limitations of the model, such as its ignoring of the geographical, cultural and linguistic differences in social media data, and not fully considering the possible impact of non-linear dynamics and external shocks (e.g., policy changes, international events) in the macroeconomic system on the predictive power of the model.

4.2.5 Outlook and Future Directions

In summary, through empirical analyses, this study confirms the potential of utilizing social media signals for macroeconomic forecasting, and also points out that there is still room for model improvement and application expansion. Future research could consider incorporating more dimensions of data (e.g., geo-location tags, multilingual content), adopting more sophisticated model fusion strategies, and developing adaptable algorithms that can effectively handle unstructured big data, in order to further explore the economic value of social media data and to enhance the model's explanatory power and forecasting accuracy. Meanwhile, interdisciplinary collaboration, combining economic theory with the latest advances in AI technology, will be key to driving in-depth research in this area.

5. CONCLUSION

Our study delves into the untapped potential of social media data within the realm of macroeconomic forecasting, accomplishing a comprehensive exploration through theoretical grounding, model development, and rigorous feasibility assessments. The key findings and contributions of this work are as follows:

We have gathered insights from information economics, behavioral economics, and big data analytics, constructing a robust theoretical scaffold that underpins the utilization of social media data for macroeconomic predictions. This interdisciplinary synthesis paves the way for innovative applications in economic forecasting.

This study was the first to apply a hybrid machine learning model to interpret macroeconomic signals embedded in a large amount of social media data. Based on the signals extracted by this model, the prediction of macroeconomic indicators is achieved, which improves the speed and accuracy of economic forecasting.

Through carefully designed experiments, we have systematically validated the causal links and predictive potency of social media indicators vis-á-vis macroeconomic variables. These multi-faceted validations underscore the immense value and untapped possibilities that social media data holds for macroeconomic forecasting, highlighting its transformative potential.

Apart from academic contributions, our research has practical implications by augmenting the analytical toolkit for policymakers, corporate strategists, and market players. It offers them a novel instrument for deciphering macroeconomic landscapes more astutely, thereby facilitating proactive decision-making and enhancing responsiveness to economic shifts.

Looking ahead, avenues for future research include: the integration of more diverse data dimensions, refining algorithms to cope with complex economic dynamics and external shocks, and advancing adaptive mechanisms that autonomously calibrate models to evolving economic contexts. Such advancements promise to further reduce the gap between digital social behaviors and macroeconomic realities, reinforcing the predictive prowess of our models and their applicability in guiding strategic policies and market maneuvers. Ultimately, this work underscores the critical importance of harnessing social media's digital power to illuminate the intricate pathways of economic forecasting.

REFERENCES

1. Aramburu, M. J., Berlanga, R., & Lanza-Cruz, I. (2023). A data quality multidimensional model for social media analysis. Business & Information Systems Engineering, 23. Doi:10.1007/s12599-023-00840-9
2. Assenmacher, D., Weber, D., Preuss, M., Valdez, A. C., Bradshaw, A., Ross, B., .et al. (2022). Benchmarking crisis in social media analytics: a solution for the data-sharing problem. Social Science Computer Review, 40(6), 1496–1522. Doi:10.1177/08944393211012268
3. Belcastro, L., Cantini, R., & Marozzo, F. (2022). Knowledge Discovery from Large Amounts of Social Media Data. Applied Sciences-Basel, 12(3), 14. Doi:10.3390/app12031209
4. Bollenbacher, J., Loynes, N., & Bryden, J. (2022). Does United Kingdom parliamentary attention follow social media posts? EPI Data Science, 11(1), 14. Doi:10.1140/epjds/s13688-022-00364-4
5. Chaki, J., Dey, N., Panigrahi, B. K., Shi, F. Q., Fong, S. J., & Sherratt, R. S. (2020). Pattern mining approaches used in social media data. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 28, 123-152. Doi:10.1142/s021848852040019x
6. Chakraborty, K., Bhattacharyya, S., & Bag, R. (2020). A Survey of sentiment analysis from social media data. IEEE Transactions on Computational Social Systems, 7(2), 450–464. Doi:10.1109/tcss.2019.2956957
7. Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007
8. Chaudhary, K., Alam, M., Al-Rakhami, M. S., & Gumaei, A. (2021). Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. Journal of Big Data, 8(1), 20. Doi:10.1186/s40537-021-00466-2
9. Chen, Y. J., & Chen, Y. M. (2022). Forecasting corporate credit ratings using big data from social media. Expert Systems with Applications, 207, 11. Doi:10.1016/j.eswa.2022.118042
10. Court, C. D., Jackson, R. W., Steele, A. J. H., Pickenpaugh, G., Járosi, P., Adder, J., & Zelek, C. (2022). Extending macroeconomic impacts forecasting for NEMS. Energy Journal, 43(4), 251–271. Doi:10.5547/01956574.43.4.ccou
11. Dendramis, Y., Kapetanios, G., & Marcellino, M. (2020). A similarity-based approach for macroeconomic forecasting. Journal of the Royal Statistical Society Series a-Statistics in Society, 183(3), 801–827. Doi:10.1111/rssa.12574
12. Dubois, E., Gruzd, A., & Jacobson, J. (2020). Journalists' use of social media to infer public opinion: the citizens' perspective. Social Science Computer Review, 38(1), 57–74. Doi:10.1177/0894439318791527
13. Fan, Y. L., Lehmann, S., & Blok, A. (2022). Extracting the interdisciplinary specialty structures in social media data-based research: A clustering-based network approach. Journal of Informetrics, 16(3), 18. Doi:10.1016/j.joi.2022.101310
14. Feuerriegel, S., & Gordon, J. (2019). News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. European Journal of Operational Research, 272(1), 162-175. Doi:10.1016/j.ejor.2018.05.068
15. Frennesson, N. F., mcquire, C., Khan, S. A., Barnett, J., & Zuccolo, L. (2023). Evaluating messaging on prenatal health behaviors using social media data: systematic review. Journal of Medical Internet Research, 25, 13. Doi:10.2196/44912
16. Guo, N., Wang, Y. Q., Jiang, H. N., Xia, X. F., & Gu, Y. (2022). TALI: An update-distribution-aware learned index for social media data. Mathematics, 10(23), 19. Doi:10.3390/math10234507
17. Hargittai, E. (2020). Potential biases in big data: omitted voices on social media. Social Science Computer Review, 38(1), 10–24. Doi:10.1177/0894439318788322
18. Hemmati, A., Arzanagh, H. M., & Rahmani, A. M. (2024). A taxonomy and survey of big data in social media. Concurrency and Computation-Practice & Experience, 36(1), 27. Doi:10.1002/cpe.7875
19. Huang, Y. Y., Gui, W. L., Jiang, Y. X., & Zhu, F. Y. (2022). Types of systemic risk and macroeconomic forecast: Evidence from China. Electronic Research Archive, 30(12), 4469–4492. Doi:10.3934/era.2022227
20. Juergens, C., & Meyer-Hess, M. F. (2021). Identification of construction areas from VHR-satellite images for macroeconomic forecasts. Remote Sensing, 13(13), 12. Doi:10.3390/rs13132618
21. Koukaras, P., Tjortjis, C., & Rousidis, D. (2020). Social Media Types: introducing a data driven taxonomy. Computing, 102(1), 295–340. Doi:10.1007/s00607-019-00739-y
22. Liu, L. B., Wang, R., Guan, W. W., Bao, S. M., Yu, H. C., Fu, X. K., & Liu, H. Q. (2022). Assessing reliability of Chinese Geotagged social media data for spatiotemporal representation of human mobility. ISPRS International Journal of Geo-Information, 11(2), 15. Doi:10.3390/ijgi11020145
23. Liu, S., & Young, S. D. (2018). A survey of social media data analysis for physical activity surveillance. Journal of Forensic and Legal Medicine, 57, 33–36. Doi:10.1016/j.jflm.2016.10.019
24. Liu, W. Z., & Cui, X. H. (2023). Improving named entity recognition for social media with data augmentation. MDPI Applied Sciences-Basel, 13(9), 11. Doi:10.3390/app13095360
25. Loaiza-Maya, R., & Smith, M. S. (2020). Real-Time macroeconomic forecasting with a Heteroscedastic Inversion copula. Journal of Business & Economic Statistics, 38(2), 470–486. Doi:10.1080/07350015.2018.1514309

26. Löchner, M., & Burghardt, D. (2023). Using hyperloglog to prevent data retention in social media streaming data analytics. ISPRS International Journal of Geo-Information, 12(2), 13. Doi:10.3390/ijgi12020060
27. Lu, X. S., Zhou, M. C., Qi, L., & Liu, H. Y. (2019). Clustering-Algorithm-Based rare-event evolution analysis via social media data. IEEE Transactions on Computational Social Systems, 6(2), 301–310. Doi:10.1109/tcss.2019.2898774
28. Mcalinn, K., Aastveit, K. A., Nakajima, J., & West, M. (2020). Multivariate bayesian predictive synthesis in macroeconomic forecasting. Journal of the American Statistical Association, 115(531), 1092–1110. Doi:10.1080/01621459.2019.1660171
29. Mcdonald, L., Malcolm, B., Ramagopalan, S., & Syrad, H. (2019). Real-world data and the patient perspective: the promise of social media? BMC Medicine, 17, 5. Doi:10.1186/s12916-018-1247-8
30. Qi, L., Li, J., Wang, Y., & Gao, X. B. (2019). Urban Observation: integration of remote sensing and social media data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(11), 4252–4264. Doi:10.1109/jstars.2019.2908515
31. Reveilhac, M., Steinmetz, S., & Morselli, D. (2022). A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion. Multimedia Tools and Applications, 36. Doi:10.1007/s11042-022-12101-0
32. Roy, K. C., Cebrian, M., & Hasan, S. (2019). Quantifying human mobility resilience to extreme events using geo-located social media data. EPI Data Science, 8, 15. Doi:10.1140/epjds/s13688-019-0196-6
33. Sadik, Z. A., Date, P. M., & Mitra, G. (2020). Forecasting crude oil futures prices using global macroeconomic news sentiment. IMA Journal of Management Mathematics, 31(2), 191–215. Doi:10.1093/imaman/dpz011
34. Sagduyu, Y. E., Grushin, A., & Shi, Y. (2018). Synthetic social media data generation. IEEE Transactions on Computational Social Systems, 5(3), 605–620. Doi:10.1109/tcss.2018.2854668
35. Shah, N., Willick, D., & Mago, V. (2022). A framework for social media data analytics using Elasticsearch and Kibana. Wireless Networks, 28(3), 1179–1187. Doi:10.1007/s11276-018-01896-2
36. Silva, T. H., & Fox, M. S. (2024). Integrating social media data: Venues, groups and activities. Expert Systems with Applications, 243, 12. Doi:10.1016/j.eswa.2023.122902
37. Taghiyeh, S., Lengacher, D. C., & Handfield, R. B. (2021). Loss rate forecasting framework based on macroeconomic changes: Application to US credit card industry. Expert Systems with Applications, 165, 18. Doi:10.1016/j.eswa.2020.113954
38. Tang, J., Tang, X. Y., & Yuan, J. S. (2018). Traffic-optimized data placement for social media. IEEE Transactions on Multimedia, 20(4), 1008–1023. Doi:10.1109/tmm.2017.2760627