Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD37151

A Review of Sentiment Analysis Technology for Social Media Text and Its Value in Consumption Prediction

Yilei Huang

The First Clinical Medical College (School of Information Engineering), Wenzhou Medical University, Wenzhou, China
18334430551@163.com

Abstract. With the rapid growth of social media, massive volumes of user-generated text data have accumulated on platforms. These unstructured expressions contain rich consumer sentiment information, which is of great significance for market research and judgment. This paper aims to clarify the technical workflow of sentiment analysis for social media text and to explore its value in consumption prediction. By sorting out the basic workflow of text preprocessing, sentiment scoring and topic identification, and by comparing the characteristics of three types of social media data sources, namely platform comments, short video danmaku and private message texts, this paper summarizes the performance of this technology in consumption prediction application scenarios such as purchase intention identification and commodity reputation evaluation. Meanwhile, this paper analyzes the impact of difficulties such as paid spam comments and ambiguous sentence recognition on the accuracy of this technology, and discusses the popularization status of low-cost sentiment analysis tools among small and medium-sized merchants. The research finds that although sentiment analysis technology can effectively mine consumption signals, social text noise and complex semantic problems still restrict prediction accuracy, and there is still large room for optimization of lightweight tools for typical merchants. In addition, in-depth exploration of multimodal data fusion and cross-platform sentiment analysis technology will become an important direction in the future. This study provides a theoretical reference and practical orientation for improving the application of social media text sentiment analysis technology in consumption prediction.

Keywords: social media, text sentiment analysis, consumption prediction, text preprocessing, multimodal analysis

1. Introduction

With the improvement of Internet infrastructure and the popularization of mobile terminals, social media has been deeply embedded in the daily life of the public. As of 2024, the number of China's Internet users has reached 1.123 billion, with an Internet penetration rate of 79.7%. Platforms such as Weibo, WeChat, Douyin and Xiaohongshu generate massive volumes of unstructured text data on a daily basis [1]. These data not only record the daily expressions of users, but also carry rich emotional tendencies and behavioral intentions in scenarios such as consumption decision-making,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD37151

brand interaction and product feedback. However, characteristics of text data including high noise, short length, semantic ambiguity and multimodal coupling render traditional manual analysis incapable of coping with the processing demand, creating an urgent need for automated and intelligent sentiment analysis technologies. How to accurately extract emotional signals from complex social texts and convert them into actionable basis for consumption prediction has become a core issue of common concern in both academia and industry.

At the commercial application level, sentiment analysis technology provides critical tools for brand reputation management, market trend prediction and user behavior insight [2]. In recent years, breakthroughs in deep learning and pre-trained language models have significantly improved the accuracy and generalization capability of sentiment classification, enabling fine-grained, aspect-level sentiment understanding [3, 4]. However, technological advancement has not fully eliminated implementation barriers: semantic pitfalls such as contextual dependency of social media text, interference from astroturfers, and irony and sarcasm still lead to unstable model performance in real-world scenarios, meanwhile, low-cost tools targeting small and medium-sized businesses have obvious shortcomings in domain adaptation, operational threshold and result interpretability, which restrict the inclusive application of this technology. Therefore, systematically sorting out the technical context, data characteristics, application scenarios and practical challenges of sentiment analysis not only helps clarify the boundaries of current research, but also provides theoretical support and practical approaches for constructing a more accurate and usable consumer prediction system.

2. Social media text sentiment analysis: technical foundations and processing workflow

2.1. Core methodological frameworks

Sentiment analysis has evolved from rule-driven to data-driven paradigms, and four main methodological approaches have emerged. Dictionary-based methods dominated the early stages. Their core logic is to perform word-level matching and polarity weighting on texts using pre-defined sentiment dictionaries (including positive words, negative words, degree adverbs, and negation terms), and then to aggregate the results into an overall sentiment orientation. The advantages of this method lie in no requirement for training data, strong interpretability and low deployment cost, however, its limitations are significant: the coverage of sentiment dictionaries is limited, making it difficult to cope with new words, internet jargon and context-dependent semantic changes, and this method cannot capture the influence of word order and syntactic structure on sentiment.

In conventional machine learning approaches, sentiment analysis is treated as a text classification task that relies on manually engineered features such as n-grams, TF-IDF and sentiment word frequency, and conducts supervised learning in conjunction with classifiers including Support Vector Machine, Naive Bayes and Random Forest [5]. Compared with lexicon-based methods, these approaches can automatically learn implicit patterns in data, with improved generalization capability, however, their feature engineering depends heavily on domain knowledge, they perform poorly on short texts and sparse features, and they struggle to capture complex semantic relationships. Deep learning methods leverage word embedding techniques (e.g., Word2Vec, GloVe) to map words into vector representations, and conduct sentiment classification with neural network architectures including CNN, RNN, LSTM and Transformer [4, 6]. Among these architectures, LSTM can effectively capture long-distance dependencies in text, while Transformer-based architectures such as RoBERTa generate high-quality word embedding representations via pre-training [4]. Aspect-Based Sentiment Analysis (ABSA) further refines the granularity of analysis: it

DOI: 10.54254/2755-2721/2026.LD37151

determines not only the overall sentiment polarity but also sentiment orientations toward specific aspects mentioned in the text, such as "screen," "battery life," and "service" [7]. Typical ABSA methods include attention-based models, graph neural network methods and generative frameworks based on pre-trained language models [7]. ABSA aligns better with real-world consumer scenarios, but it features higher task complexity, relies on high-quality aspect term annotation data, and still confronts challenges when aspect terms are implicit or involve cross-sentence coreference.

2.2. Standardized processing pipeline

The engineering implementation of sentiment analysis generally follows a three-stage standardized workflow: text preprocessing serves as the fundamental component, whose objective is to convert raw social media text into a standardized format acceptable as model input. The core operations include: removing non-text noise such as HTML tags, URLs, and emojis, word segmentation and part-of-speech tagging (word segmentation is required for Chinese, while English is split by spaces), stopword filtering (retaining emotion-critical functional words such as negative words and degree adverbs), text standardization (traditional-simplified Chinese conversion, case unification, and typo correction), and augmentation strategies for short texts (e.g., synonym substitution and context concatenation). The quality of preprocessing directly affects the performance of subsequent models, especially given the high noise and informal expression characteristics of social media text, thus cleaning rules need to be customized in combination with domain knowledge.

Sentiment polarity scoring constitutes the core step, whose objective is to assign continuous or discrete sentiment intensity values to the given text or aspect. The mainstream methodologies can be categorized as follows: Lexicon weighting methods (e.g., SO-PMI, SentiStrength) output sentiment scores within the interval [-1, 1], machine learning/deep learning classifiers output class probabilities (e.g., positive with probability 0.85), and regression models directly predict the scalar of sentiment intensity. In consumption prediction tasks, continuous scores usually need to be discretized into operational labels such as "strongly positive," "weakly positive," "neutral," "weakly negative," and "strongly negative" based on business-specific thresholds, and then aggregated over time to capture trends.

Topic semantic recognition serves as the value-enhancement step, whose objective is to anchor sentiment signals to specific consumption objects or behavioral intentions. Commonly adopted methods include: automatic clustering via topic models such as LDA and BERTopic, aspect-based sentiment analysis (ABSA) based on aspect term extraction, and entity-relation-sentiment triple extraction integrated with knowledge graphs. This step upgrades sentiment analysis from "emotion detection" to "semantic understanding", and supports high-level applications including purchase intention identification, word-of-mouth attribution, and risk early warning.

3. Typical data types and characteristics of social media text

3.1. Review data from e-commerce platforms

E-commerce reviews represent the most mature data source for consumer sentiment analysis, with typical characteristics including long text, multi-dimensional evaluation, and relatively high structuralization. Users usually elaborate on multiple attributes of a product, such as appearance, performance, logistics and after-sales service, accompanied by descriptions of specific usage scenarios and comparative references, which provide ideal corpora for aspect-based sentiment analysis (ABSA) [7]. However, reviews also contain a large amount of templated positive

DOI: 10.54254/2755-2721/2026.LD37151

comments, fraudulent brushing orders and emotional venting, which requires noise filtering via multi-dimensional features including text length, repetition rate and user behavior. The analytical value of this data type lies in its direct correlation with sales data, which require noise filtering via multi-dimensional features such as text length, repetition rate, and user behavior.

3.2. Short video danmaku data

Danmaku (bullet comments) is a unique real-time commenting feature exclusive to video platforms, characterized by short text, strong timeliness, high interactivity, and heavy reliance on context. A single danmaku usually contains only several to a dozen of Chinese characters, and mostly reflects viewers' immediate reactions (such as " dying of laughter", "emotionally broken", "product recommendation", and "product warning"). Its emotional expression is highly condensed and relies heavily on the video content as well as the context of preceding and subsequent danmaku [1, 8]. The main difficulties of danmaku analysis lie in semantic fragmentation and multimodal coupling, while its unique advantage lies in capturing the instantaneous outbreak of consumer sentiment and group resonance, which makes it suitable for real-time public opinion monitoring and sentiment popularity tracking in scenarios such as new product releases, live-streaming commerce and hot events.

3.3. Private messages on social platforms

Private messages refer to private communications among users in one-to-one or group scenarios, characterized by high personalization, strong intention orientation and low publicity. The content of private messages often contains authentic purchase inquiries, usage confusion, after-sales demands and even competitor product comparisons, with more natural emotional expression, richer details, and less susceptibility to social desirability bias. The analytical value of private messages lies in their ability to support the construction of in-depth user profiles and to help identify potential user needs, churn risks, and word-of-mouth communication nodes. However, the acquisition of private message data is strictly restricted by privacy policies, and such data usually requires user authorization or can only be obtained indirectly through customer service systems, leading to limitations in analysis scale and representativeness.

4. Application scenarios of sentiment analysis technology in consumption prediction

4.1. Identification of users' purchase intention

Sentiment analysis can help identify users' purchase intentions through several analytical mechanisms. First, the sentiment intensity threshold serves as a critical judgment basis: strong positive sentiments (such as "such a pleasant surprise" and "a must-buy") usually correspond to a high conversion probability, while neutral or weak positive sentiments may only indicate that users are in the information collection stage. Second, the frequency with which core product attributes are mentioned can also reflect users' intentions. If users repeatedly mention key decision-making attributes such as "battery life" and "cost-performance ratio" and express positive sentiments towards them, it indicates that they are in the in-depth evaluation stage. In addition, the proportion of decision-related keywords also has indicative significance. When words such as "place an order", "already purchased" and "recommend to friends" co-occur with positive sentiments, they often constitute a strong purchase signal [2, 9]. In commercial scenarios, this logic can be widely applied to mining high-potential users from e-commerce reviews for precision marketing, or distinguishing

DOI: 10.54254/2755-2721/2026.LD37151

authentic product recommendations from native soft advertisements through joint sentiment-attribute analysis of product promotion content (such as Xiaohongshu notes). Meanwhile, the temporal alignment of words such as "want it" and "link" in live streaming danmaku with emotional peaks can also assist streamers in adjusting their presentation scripts in real time.

4.2. Product public opinion evaluation and risk early warning

Public opinion evaluation relies primarily on sentiment trend monitoring and early warning mechanisms for abnormal fluctuations. On the one hand, by tracking the time-series sentiment curve of a product after its launch, and observing changes in sentiment mean and variance, continuous declines or sharp fluctuations often indicate potential problems. On the other hand, combining aspect-based sentiment analysis (ABSA) to conduct negative sentiment attribution enables the localization of aspects with concentrated negative feedback (e.g., "battery swelling", "unresponsive customer service"), thereby distinguishing between individual complaints and systematic defects. In addition, applying Statistical Process Control (SPC) or the Isolation Forest algorithm for anomaly detection can identify mutation points in sentiment distribution and trigger manual review. In business practice, this mechanism can be applied to rapid 72-hour public opinion reports after new product launches to guide public relations responses, to identify own shortcomings through comparative public opinion analysis with competing products, and even to issue early warnings for surging negative sentiment during major promotion periods to prevent public opinion from escalating. This application shifts sentiment analysis from "post-event summarization" to "in-event intervention", significantly improving the timeliness of risk management and control.

5. Practical challenges of technology application and current status of deployment tools

5.1. Core technical challenges

Currently, sentiment analysis for consumption prediction faces two prominent technical challenges, both of which directly impair business decision-making:

First, fraudulent reviews from malicious actors distort word-of-mouth perceptions and lead to misjudgment by merchants. Professional review fraud rings systematically contaminate review data through template-based positive reviews, bulk account registration, and simulated real user behaviors [8, 10]. Traditional sentiment models struggle to distinguish between genuine users and bots, and often misclassify fraudulent positive reviews as authentic positive word-of-mouth. This causes merchants to overestimate product acceptance and incorrectly increase inventory or marketing investment. Conversely, malicious negative reviews may also be misjudged as authentic negative feedback, triggering excessive public relations responses or even unwarranted product removal. Such noise not only reduces model accuracy, but also undermines the decision-making foundation of consumption prediction.

Second, ambiguous utterances are misclassified by algorithms, which impairs accurate word-of-mouth analysis. A large number of social media texts contain rhetorical devices such as irony, sarcasm and metaphor (for example, "The quality is just 'amazing'" and "Customer service is so 'thoughtful' that it took three days to reply"), whose literal sentiment contradicts the actual intention. Although existing models achieve good performance on standard test sets, they still frequently make misclassifications in real social contexts: they may recognize sarcasm as positive sentiment, or classify subtle criticism as neutral. Such errors are particularly dangerous in word-of-mouth analysis: on the one hand, they cover up real problems and delay product improvement, on the other

DOI: 10.54254/2755-2721/2026.LD37151

hand, if sarcasm is misjudged as positive sentiment and used in marketing campaigns, it may trigger user resentment and a crisis of trust.

5.2. The popularization status of low-cost tools for small and medium-sized merchants

The popularization of sentiment analysis tools among small and medium-sized merchants is characterized by "high awareness, low application, and limited effectiveness". In terms of penetration, although leading e-commerce platforms (such as Taobao and JD.com) have built-in basic sentiment analysis functions, and third-party SaaS tools (such as Weimob and Youzan) also provide sentiment tagging services, most small and medium-sized merchants still rely on manual sampling review, and the utilization rate of automated analysis is less than 30%. In terms of core demands, merchants attach the highest priority to three characteristics: "zero operational threshold", "interpretable results" and "compatibility with existing business systems", rather than purely model accuracy. They expect tools to directly output business conclusions such as "negative comment early warning", "hit product potential identification" and "competitor comparison", instead of raw sentiment scores.

Nevertheless, currently available tools still have obvious shortcomings: First, insufficient vocabulary coverage. General dictionaries can hardly cover vertical industry terms (such as those in medical aesthetics, pet food and Hanfu) and emerging internet jargon,leading to missed or misclassified results. Second, poor domain adaptability. Most tools are trained on general corpora, leading to a sharp performance drop in specific industries such as maternal and infant products, digital electronics and food, while also lacking fine-tuning interfaces. Third, mismatch with merchants' digital literacy. Most small and medium-sized merchants lack the capability to interpret data, and feel confused when facing abstract indicators such as "average sentiment score of 0.65". Meanwhile, existing tools lack conversion into business language and attribution explanation, which results in the problem of "data availability without actionable insights".

6. Conclusion

Sentiment analysis technology has formed a complete methodological system ranging from lexicon-based methods to pre-trained models. Aspect-level analysis and standardized processing procedures are key foundations for using this technology to support consumption prediction, yet technological advances have not fully addressed the challenges posed by the semantic complexity and noise in social texts. Different social data sources feature differentiated characteristics and application adaptability: e-commerce reviews are suitable for aspect-level word-of-mouth attribution, danmaku are applicable to real-time emotion monitoring, and private messages are ideal for constructing in-depth user profiles. It is thus necessary to select appropriate data sources and analysis granularity based on specific business objectives. The effectiveness of consumption prediction is highly dependent on the combination of technical accuracy and business interpretability. At present, technical challenges such as opinion spam interference and ambiguous misjudgment, as well as the shortcomings of existing tools in terms of domain adaptability and operational threshold, remain the main bottlenecks restricting its large-scale implementation.

Existing research still has certain limitations. In terms of data acquisition, data such as private messages and danmaku are restricted by platform privacy policies and API permissions. Most empirical studies rely on public datasets or simulated data, which creates a gap between research data and real business scenarios. In terms of evaluation criteria, existing research mostly adopts general metrics such as F1-score and accuracy, and lacks business-oriented indicators specifically

DOI: 10.54254/2755-2721/2026.LD37151

for consumption prediction (such as monetary loss caused by misjudgment, early warning lead time, etc.), making it difficult to quantify the actual practical value of the technology. In addition, cross-platform generalization verification is insufficient: most models are only validated on a single platform (such as Weibo and Taobao), and cross-platform migration performance and domain adaptation mechanisms have not been fully explored.

Future research and practice can be developed in three directions: In terms of technology, researchers should develop robust sentiment analysis models resistant to noise and opinion spam, integrate user behavior, temporal patterns and multimodal signals to construct a multi-dimensional verification mechanism, and explore the potential of large language models in zero-shot aspect-level analysis and irony recognition. In terms of implementation and application, researchers should develop "business-oriented" sentiment analysis tools for small and medium-sized merchants, with built-in industry knowledge bases, automatic attribution interpretation and risk level visualization, and provide digital literacy training and decision support templates to promote the transformation of the technology from "available" to "easy-to-use". In terms of ethical norms, an ethical framework for social text sentiment analysis should be established, clarifying data usage boundaries, user informed consent mechanisms and algorithmic bias auditing requirements, preventing the abuse of emotion monitoring and privacy infringement, and ensuring that technological application conforms to social values and regulatory requirements.

References

[1] Tang, L. Y., Xiong, C. C., Wang, Y., et al. (2021). A review of short text sentiment analysis based on deep learning. *Journal of Frontiers of Computer Science and Technology, 15*(5), 794–811.

[2] Yan, W. X. (2024). Research on consumer behavior in the context of the rise of social media. *Modern Marketing*, (5), 54–56.

[3] Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific and Technology Research, 9*(2), 601–609.

[4] Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network. arXiv preprint. https://doi.org/10.1109/access.2022.3152828

[5] Chen, Q., Che, M. M., Guo, X., et al. (2021). A text sentiment classification method based on recurrent convolutional attention model. *Computer Science, 48*(2), 245–249.

[6] Peng, Z. W., Zhu, X. F., & Guo, J. F. (2022). Label-guided dual attention deep neural network model. *Pattern Recognition and Artificial Intelligence, 35*(2), 175–184. https://doi.org/10.16451/j.cnki.issn1003-6059.202202008

[7] Wang, H. S., Li, J. Z., & Zeng, B. Q. (2023). A review of aspect-level sentiment analysis methods. *Software Guide, 22*(9), 1–8.

[8] Al-Tameemi, I. K. S., Feizi-Derakhshi, M. R., Pashazadeh, S., & Asadpour, M. (2024). A comprehensive review of visual–textual sentiment analysis from social media networks. *Journal of Computational Social Science, 7*(3), 2767–2838.

[9] Gupta, S., & Sandhane, R. (2022). Use of sentiment analysis in social media campaign design and analysis. *Cardiometry, 22*(2022), 351–363.

[10] Yadav, J. (2023). Sentiment analysis on social media. *arXiv preprint*. https://doi.org/10.32388/yf9x04