Dean&Francis
ISSN 2959-6157

# Research on NLP Techniques for Analyzing Social Media User Behavior

**Yufan Shi**

Taicang Campus, Xi'an Jiaotong-Liverpool University, Taicang, Suzhou, Jiangsu, China
Yufan.Shi23@student.xjtlu.edu.cn

**Abstract:**

This paper presents a systematic review of the technical framework and notable advancements in natural language processing (NLP) technology for analyzing social media user behaviorIn the study of text feature extraction, the initial focus is on the evolution of semantic methodologies, transitioning from bag-of-words techniques to more sophisticated word vector models. It compares Word2Vec, GloVe, and FastText in terms of how well they can convey semantics, how much processing power they need, and how well they can handle words that aren't in their lexicon. The study evaluates the limitations of coarse-grained sentiment analysis for sentiment feature extraction and examines how fine-grained methodologies, like Aspect-Based Sentiment Analysis and Emotion Cause Extraction, rectify these deficiencies through specifically delineated tasks and customised model architectures. The research analyses performance disparities in behavioural categorisation across traditional shallow models, such as SVM and LSTM, highlighting their inherent limitations, and emphasises BERT's advantages in utilising bidirectional context to improve classification results.This systematic technical approach includes text feature extraction, sentiment analysis, and behavioral categorization, offering a reference for analyzing social media user behavior, a research methodology for related studies, and a foundation for multiple field integration.

**Keywords:** NLP; sentiment analysis; behavior analysis.

## 1. Introduction

Currently, social media has become an important channel for expressing opinions and emotions. Globally, approximately 5.56 billion people were connected to the internet, with social media penetration reaching 5.24 billion individuals as of February 2025, representing 67.9% and 63.9% of the global population, respectively [1]. Extensive user engagement on social platforms generates substantial data volumes through daily interactions. As an illustrative example, Twitter (X) sustains about 611 million monthly active users, with monthly user engagement averaging 15.67 hours [2]. Additionally, daily posts alone can

amount to 500 million in 2024 [3]. These datasets possess substantial value: companies rely on data analysis to improve recommendation systems and user retention; industries use online reviews for preference mining and product marketing [4]; and government agencies utilize opinion monitoring in their crisis management and event prediction efforts [5]. However, due to the large volume of data produced by social media, traditional matching methods are inadequate to keep up with the rapid evolution of internet terminology, while manual annotation becomes too expensive. The use of natural language processing (NLP) techniques offers a new method to resolve the above issues by promoting the improvement of user behavior understanding through automated feature extraction and sentiment analysis.

NLP techniques involve some crucial branches, including information extraction, machine translation, dialogue systems, and sentiment analysis. Of these, sentiment analysis is the most appropriate one for social media user behavior analysis because it aims at automatically recognizing subjective emotions, opinions, attitudes, and their intensity in text. This kind of emotional information not only prevails extensively in social data but also shows users' preferences directly, making it the core approach for user behavior analysis.

The purpose of this research is to provide a detailed investigation of the technological advancements in this field, with special focus on the core methods utilized in sentiment analysis. Systematically, it will study the technical underpinnings that enable feature extraction, classification, and sentiment analysis; analyze the effectiveness of models from different eras in performing corresponding tasks; and integrate the characteristics of the corresponding models for enabling research efforts in the future.

## 2. Task Decomposition

### 2.1 Text Feature Extraction

Machines cannot analyze social media text directly and require preprocessing through text feature extraction methods. The process maps unstructured characters into structured computational features, mainly with the purpose of translating linguistic symbols into mathematical vector representations while preserving semantic, syntactic, and contextual information. The processing workflow comprises several key steps: text preprocessing to cleanse and standardize data for consistency; vocabulary normalization to unify text formats and reduce variability; feature construction to identify suitable text representations; vectorization to transform text into numerical vectors for analysis; and feature optimization to enhance feature quality and relevance, thus enhancing model performance.

#### 2.1.1 Semantic features extraction

Semantic Feature Extraction emphasizes the identification of the user's main message. For instance, in „I need help with something" and „I can help with something," both statements include the terms „help" and „something," yet the former conveys a request for „help," while the latter extends an offer of „help." The model must initially grasp the objective semantic representation of the text to ensure accurate understanding of user expressions.

Initial text feature extraction primarily depended on superficial statistical features, like bag-of-words models and TF-IDF. These approaches are computationally efficient and economical, making them suitable to handle small-scale data. Nevertheless, they are characterized by exceedingly large feature dimensions with minimal non-zero element ratios, and they overlook part-of-speech and semantic information. Subsequent word vector methods represent words in low-dimensional, dense vector spaces to effectively capture semantic relationships among them. Table 1 presents a comparison of three representative models for the implementation of word vectors, emphasizing the distinct characteristics of each model.

**Table 1. Comparison of Word2Vec, GloVe, and FastText**

| Method | Suitable Corpus Size | Advantages | Disadvantages |
|---|---|---|---|
| Word2Vec | Medium size | Captures semantic and syntactic relationships; fast computing | One meaning per word; Ignores global statistics; Poor handling of rare words |
| GloVe | Large scale | Improve global statistical data use; Better semantic and syntactic stability | Needs fixed word vectors; Weak at representing unregistered words |
| FastText | More versatile | Supports unseen words; suitable for complex morphology languages | Insufficient optimization for short or high-frequency words; subword combinations could create noise |

ISSN 2959-6157

Among these methods, Word2Vec was proposed by Mikolov et al. [6]. Word2Vec encompasses two new architectures: CBOW (Continuous Bag of Words) and Skip-gram. The former predicts the central word based on the context, while the latter forecasts the context based on the central word. Although the models predict in different directions, both utilize word co-occurrence patterns to derive word vectors, which exhibit linear algebraic properties facilitating analogical reasoning; for example, the operation „bigger - big + small“ directly produces „smaller“ within the semantic space. Experimental findings indicate that these novel models substantially surpass conventional methods like RNNLM and NNLM regarding semantic precision [6]. The CBOW model exhibits high syntactic accuracy but constrained semantic representation, whereas Skip-gram displays more balanced proficiency. Moreover, Skip-gram typically demonstrates enhanced overall learning efficacy relative to CBOW. Nevertheless, the enhanced efficacy of Skip-gram necessitates increased computational resources, thus resource efficiency must be meticulously assessed when examining extensive corpora.

GloVe overcomes the constraint of depending exclusively on local co-occurrence statistics within sliding windows by utilizing a word co-occurrence probability matrix, thus attaining global statistical integration that more effectively captures semantic patterns [7]. To evaluate the semantic relationship between two words i and j, it is essential to analyze their co-occurrence probability ratios with arbitrary probe words k (i.e., P(k|i)/P(k|j)), instead of considering their individual probabilities independently. This model surpasses simple analysis of co-occurrence frequency between word pairs (i,j) by utilizing multiple probe words k to assess the relative association strength of k in the context of i compared to j. For example, utilizing „gas“ as a probe word to compare „rocket“ and „fish,“ if the co-occurrence probability ratio P(gas|rocket)/P(gas|fish) markedly exceeds 1, it signifies that „rocket“ possesses a robust semantic connection to fuel characteristics, whereas „fish“ does not.

FastText addresses the out-of-vocabulary (OOV) issue by incorporating sub-word information, dividing down words into sets of character n-grams Consequently, even when previously unseen words are encountered during training, the sub-words derived from decomposition are likely to have been present in the training corpus [8]. The vector of the new term can be derived by amalgamating established segment vectors. For instance, „unbelievable“ can be segmented into character n-grams such as „un“, „unb“, „nbe“, „bel“, „eli“, „lie“, „iev“, „eve“, „vea“, „eab“, „abl“, „ble“, where significant components like „un“ (denoting negation), „bel“/“lie“/“iev“ (elements of „believe“), and „abl“/“ble“ (components of „able“) enable the model to grasp the composite meaning of „not able to be believed“ through the amalgamation of these subword units.

Liu et al. [9] and Modha and Majumder [10] found that FastText consistently outperforms Word2Vec and GloVe on social media datasets, including Facebook and Twitter, in tasks such as emotion prediction and text classification. Due to the prevalence of spelling errors, abbreviations, mixed languages, and extensive slang in social media texts, FastText possesses a distinct advantage in addressing out-of-vocabulary issues and is more adept for such content.

### 2.1.2 Emotion feature extraction sub heading

In addition to semantic attributes, behavior classification necessitates the direct acquisition of emotional signals that elucidate „why users articulate themselves in this manner,“ enhancing content comprehension with a motivational aspect. Examples such as „I can‘t believe this works“ and „I can‘t believe this happened“ both articulate shock; nevertheless, the first usually conveys a sense of happiness, whilst the latter frequently denotes sadness. People must capture emotional signals in addition to semantic meaning by constructing affective variables to assist models in recognizing user expressions that share similar semantic topics but differ in behavioral intentions.

Conventional sentiment analysis operates at the document or sentence level as coarse-grained sentiment analysis (CSA), which analyzes the overall sentiment polarity (positive, negative, or neutral) of a complete text segment. However, user expressions usually demonstrate complexity and multidimensionality. As a result, Fine-grained Sentiment Analysis (FSA) has emerged. FSA includes various subtasks, notably Aspect-Based Sentiment Analysis (ABSA) [11] and Emotion Cause Extraction (ECE), which are frequently examined. Table 2 illustrates the distinct outputs of fine-grained and coarse-grained analyses for the same example:

**Table 2. Comparison between CSA and FSA**

| Example: The store’s signature dessert is unexpectedly tasty, but the wait is simply too lengthy. | |
|---|---|
| TASK | OUTPUT |
| CSA | Overall sentiment of the sentence: negative |

| | | |
|---|---|---|
| FSA | ABSA | Aspect(dessert): sentiment (positive)<br>Aspect(wait): sentiment(negative) |
| | ECE | Emotion(joy): emotion-causing text fragment (“unexpectedly tasty”)<br>Emotion(anger): emotion-causing text fragment (“imply too lengthy”) |

ABSA aims to discern particular aspects from text and evaluate their associated polarity of sentiment, effectively distinguishing sentiment orientations towards various aspects within the same text. This process depends on the cooperative functioning of four sentiment components. The Aspect Category delineates a predefined framework of abstract concepts; the Aspect Term identifies specific entities explicitly mentioned in the text; the Opinion Term extracts vocabulary expressing sentiment towards that entity; and the Sentiment Polarity compiles a quantified judgment of positive, negative, or neutral, thus finalizing the systematic extraction of a detailed sentiment structure. In accordance with the example presented in Table 2: Identify Aspects („dessert“, „wait“) → Extract Opinions („tasty“, „too lengthy“) → Assign Sentiment (Positive, Negative) Assign to Categories (Food Quality, Service Efficiency) → Output Structured Sentiment {dessert: positive, wait: negative}.

To efficiently represent local dependencies and long-distance relationships in text, researchers frequently employ model fusion strategies, utilizing BERT/RoBERTa as robust text encoders and subsequently inputting their contextual representations into alternative network architectures. Prevalent fusion models comprise: BERT/RoBERTa + BiLSTM/BiGRU, BERT/RoBERTa + CNN, BERT/RoBERTa + GCN/GAT... Simultaneously, attention mechanisms are extensively employed to enhance associations between aspect terms and pertinent context. For example, Cheruku et al. proposed an integration of an enhanced RoBERTa with RNNs, attaining 84.6% accuracy on Twitter review datasets [12]. Feng et al. conducted experiments that merged convolutional neural networks with attention models for text sentiment analysis, elevating accuracy to 87% [13].

ECE, a subtask of fine-grained sentiment analysis, concentrates on identifying and extracting text segments (cause spans) that elicit recognized emotions, effectively addressing the question of „why this emotion arises.“ Table 3 illustrates the progression of ECE methodologies:

Lee et al.initially conceptualized emotion cause extraction as word-level sequence labeling, created a small-scale dataset with boundary annotations for emotion and cause words derived from the „Academia Sinica Balanced Chinese Corpus,“ and developed rule-based systems by synthesizing linguistic indicators, including causal connectives and emotional verbs [14].

Gui et al. advanced ECE from word-level to clause-level binary classification to ascertain whether each clause serves as an emotion cause [15]. A comprehensive Chinese corpus, the Sina News Corpus, was developed from „Sina City News,“ and neural network models (CNN/LSTM) were initially employed in ECE tasks, demonstrating the efficacy of deep learning in these applications.

Xia and Ding proposed a method for the direct extraction of emotion-cause pairs, implementing an Inter-EC mechanism and constraint-based pair filtering to address the limitations of Emotion-Cause Extraction (ECE) that necessitates pre-annotated emotions [16].

**Table 3. Evolution of ECE Methods**

| Method | Word-level Sequence Labeling | Clause-level Binary Classification | Emotion-Cause Pair Extraction |
|---|---|---|---|
| Contributions | Pioneered ECE task definition;<br>Constructed first Chinese emotion cause corpus;<br>Proposed rule-driven methods | Restructured task definition;<br>Released first open-source benchmark dataset;<br>Verified deep learning effectiveness | Directly extracted emotion-cause pairs jointly;<br>Designed two-stage model |

RoBERTa demonstrates outstanding efficiency and is often utilized as a baseline for ECE. To improve performance, researchers frequently enhance RoBERTa by combining it with BiGRU, attention mechanisms, and additional models to enhance its capacity to discern causal relationships in lengthy texts. Data from training and evaluating various models on the Google-developed GoEmotions dataset indicates that traditional machine learning models exhibit markedly inferior performance compared to NLP models, with the exception of GPT [17]. RoBERTa surpasses the other two NLP models in all evaluation metrics, illustrating its strong proficiency in fine-grained emotion classification tasks [17].

## 2.2 Behavioral Classification

After collecting the aforementioned multidimensional fea-

ISSN 2959-6157

tures, the challenge becomes devising a classifier capable of precisely differentiating user behavior types.

Conventional approaches generally rely on manual feature engineering complementing autonomous classification algorithms. For instance, superficial models like Support Vector Machines (SVM) can rapidly execute keyword classification for overt behaviors, whereas deep sequence models such as Long Short-Term Memory networks (LSTM) can proficiently discern user intentions. Nonetheless, these methods possess specific limitations in identifying implicit behaviors and managing intricate contextual information.

Subsequently, researchers found that pre-training the model on a substantial corpus of unlabeled text prior to executing specific tasks significantly enhanced outcomes. However, these models are unidirectional, allowing them to perceive context solely from the current word in one direction, either leftward or rightward. Consequently, only fifty percent of the information is accessible, resulting in the acquisition of only partial information, which hampers thorough comprehension.

Bidirectional Encoder Representations from Transformers (BERT) provides an effective solution to this problem [18]. This phenomenon stems from BERT‘s requirement for the Transformer architecture to simultaneously assess relationships between preceding and subsequent words in context. BERT offers two pre-training tasks: the Masked Language Model, similar to cloze tests, and Next Sentence Prediction, which evaluates the logical progression of sentence B following sentence A. The two tasks attempt to enhance the model‘s effectiveness in natural language understanding. Both tasks achieve bidirectional pre-training, and BERT‘s multi-layered architecture with a sequential layer-by-layer design facilitates the extraction of more complex features, effectively capturing long-range dependencies and intricate syntactic structures. Furthermore, BERT demonstrates significant advantages in text feature extraction. Wei et al. found that BERT outperformed SVM and LSTM in accuracy, recall, and F1 scores using the same training dataset, indicating superior classification performance [19].

## 3. Conclusion

This paper presents a systematic review of the complete technical framework of natural language processing technologies employed in the analysis of social media user behavior, encompassing semantic feature extraction, sentiment feature extraction, and behavior classification. It analyzes the development and efficacy of techniques including bag-of-words models, word embeddings (Word-2Vec/GloVe/FastText), fine-grained sentiment analysis (ABSA/ECE), and deep learning classification models (SVM/LSTM/BERT). This study primarily concentrates on text content analysis and does not extensively explore interactive behaviors like likes and shares, nor does it consider multimodal data such as images and emojis. Furthermore, researcher persist in encountering obstacles in this research domain. At the data level, cross-platform data heterogeneity complicates information sharing, while issues with dialect processing and sparse annotation are prevalent. Theoretically, people encounter difficulties in the dynamic evolution of language and multimodal sentiment expression, especially when non-textual elements significantly contribute to emotional conveyance, as conventional text analysis models exhibit limited explanatory capacity in these instances. Future research ought to establish an analytical framework that integrates multimodal data, interdisciplinary viewpoints including social psychology, and enhances model adaptability and interpretability regarding complex emotional drivers. This will propel the extensive utilization of social media analysis in domains like public opinion assessment and tailored recommendations, leading to a paradigm shift from isolated text analysis to holistic behavioral comprehension.

## References

[1] DataReportal, Meltwater, We Are Social. Number of internet and social media users worldwide as of February 2025. Feb. 5, 2025. Available at: https://www.statista.com/statistics/617136/digital-population-worldwide/

[2] Singh S. How many people use social media (2025 usage stats). DemandSage, Aug. 1, 2025. Available at: https://www.demandsage.com/social-media-users/

[3] Singh S. Twitter (X) user statistics 2024. DemandSage. Available at: https://www.demandsage.com/twitter-statistics

[4] Dhar S, Bose I. Corporate users’ attachment to social networking sites: Examining the role of social capital and perceived benefits. Information Systems Frontiers, 2023, 25: 1197–1217.

[5] Wang N, Lv X, Sun S, et al. Research on the effect of government media and users’ emotional experience based on LSTM deep neural network. Neural Computing & Applications, 2022, 34: 12505–12516.

[6] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[7] Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532–1543.

[8] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification. In Proceedings of the 15th

Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017: 427–431.
[9] Liu F, Hou K, Dong Y. Deep parallel contextual analysis framework based emotion prediction in community wellness communications on social media. Heliyon, 2024.
[10] Modha S, Majumder P. An empirical evaluation of text representation schemes on multilingual social web to filter textual aggression. arXiv preprint arXiv:1904.08770, 2019.
[11] Pontiki M, Galanis D, Papageorgiou H, et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In Proc. 10th Int. Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 2016: 19–30.
[12] Cheruku R, Hussain K, Kavati I, et al. Sentiment classification with modified RoBERTa and recurrent neural networks. Journal of Reliable Intelligent Environments, 2023, 9(1): 1–12.
[13] Feng X, Zhang Z, Shi J. Text sentiment analysis based on convolutional neural network and attention model. Journal of Computer Applications Research, 2018, 35(5): 1434–1436.
[14] Lee SYM, Chen Y, Huang CR. A text-driven rule-based system for emotion cause detection. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, June: 45–53.
[15] Gui L, Wu D, Xu R, Lu Q, Zhou Y. Event-driven emotion cause extraction with corpus construction. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), 2016: 1639–1649.
[16] Xia R, Ding Z, Liu X. Emotion-cause pair extraction: A new task to emotion analysis in texts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), 2019: 1003–1012.
[17] Zhang X, Qi X, Teng Z. Performance evaluation of Reddit comments using machine learning and natural language processing methods in sentiment analysis. arXiv, 2024.
[18] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, 2019: 4171–4186.
[19] Wei L, Li P, Zhang G, Cheng Q. Recognition of lexical functions in academic texts: Automatic classification of keywords based on BERT vectorization. Journal of the China Society for Scientific and Technical Information, 2020, 39(12): 1320–1329.