Exploring Public Attitudes toward Artificial Intelligence through Social Media Text Analysis

Xinye Zhou

Business School, Geely University, Chengdu, China
Email: 814408118@qq.com
Manuscript received September 7, 2026; accepted October 20, 2026; published November 12, 2026.

***Abstract*—Sentiment analysis of these texts helps identify the public's diverse attitudes toward AI. This study focuses on AI-related texts from social media. After text cleaning, manual annotation, and dataset partitioning, a three-class dataset (positive, neutral, and negative) was constructed. Text features were extracted using methods such as TF-IDF and BERT, and classification experiments were conducted using Logistic Regression and SVM. The classification performance of different models was compared using 424 test texts. The results show that the accuracy of the TF-IDF+Logistic Regression, TF-IDF+SVM, BERT+SVM, and BERT+Logistic Regression models were 78.77%, 81.84%, 85.14%, and 86.56%, respectively. Among these, the BERT+Logistic Regression model achieved the best overall classification results, with a Macro-F1 score of 86.32%. Further analysis revealed that the classification performance for neutral text was lower than that for positive and negative text, with misclassifications primarily occurring between neutral text and other categories. The study demonstrates that incorporating contextual semantic features can enhance the ability to identify attitudes in social media text and also highlights the complexity of public attitudes toward artificial intelligence.**

***Keywords*—social media, artificial intelligence, sentiment analysis, BERT, text classification**

I. Introduction

With the continuous advancement of the digital economy and digital transformation, artificial intelligence has gradually become a key technology for the power industry to enhance supply chain management and customer service capabilities. Currently, technologies such as intelligent customer service, automated outbound calling, speech recognition, intelligent voice interaction, multimodal knowledge bases, and intelligent work order quality inspection are increasingly being applied to supply chain customer service, playing a role in reducing repetitive manual work, improving response efficiency, and enhancing the user experience [1]. However, as business scenarios and customer demands continue to expand, traditional customer service still faces challenges such as high labor costs, insufficient response efficiency, inconsistent service quality, and imprecise identification of user needs. Therefore, how to leverage artificial intelligence to further enhance the level of service intelligence has become a technical issue worthy of attention.

At the same time, as the public's exposure to artificial intelligence continues to grow, users have generated a vast amount of AI-related comments and discussions on social media. These sentences not only express positive opinions on the effectiveness, convenience and interaction of intelligent services but also show the public's worries about the accuracy, privacy and security. Due to the huge amount and complexity of the social media texts, the traditional manual examination cannot process them efficiently [2]. The model is assessed by using classification criteria and error analysis, which offers a technical basis for the study of user feedback in intelligent customer service.

II. BERT-based Social Media Text Processing and Public Attitude Classification Method

A. Social Media Text Data Collection and Preprocessing

Social media messages possess an unordered feature, containing various types of expressions and varying lengths of text; dealing with them directly may be affected by unnecessary characters, repeated texts and abnormal information [3]. Hence, in this paper, the social media messages related to AI are firstly purified and standardized. The preprocessing includes the removal of duplicate texts, the rejection of invalid texts, the treatment of special characters, the normalization of texts and the determination of sentiments by hand. While ensuring that the original semantic meaning remains essentially unchanged, the text is converted into standardized data suitable for processing by the BERT model.

Text preprocessing can be represented as:

$$D' = \{T_i \mid T_i \in D, C(T_i) = 1\} \tag{1}$$

where D represents the original set of social media texts, T represents the text at position i , $C(T_i)$ represents the text validity judgment function, which returns 1 when the text meets data quality requirements and 0 otherwise, and D' represents the set of valid texts after filtering.

After text cleaning is complete, manual annotation is performed based on the primary attitude expressed in the text, categorizing public attitudes toward artificial intelligence into three types: positive, neutral, and negative. To ensure effective training of the subsequent classification model, it is important to ensure that each category has a sufficient number of samples and to divide the data into training and testing sets.

B. BERT-based Semantic Feature Extraction from Text

Traditional text representation methods typically convert text into words or word-frequency features; however, social media text contains a large amount of context-dependence and semantic variation, making it difficult to fully describe the meaning of the text by relying solely on word frequency. To improve the semantic representation capability of text, this paper adopts the BERT (Bidirectional Encoder Representations from Transformers) model to extract features from preprocessed social media text.

BERT is based on the Transformer encoder architecture

and learns semantic relationships between different words in a text through bidirectional contextual information [4]. For the input text, tokenization is first performed to convert the text into a sequence of tokens recognizable by the model, with special tokens added. Subsequently, the tokens are converted into vector representations via the embedding layer, and context-relevant semantic features are obtained through multiple layers of the Transformer encoder.

For the input sequence:

$$X = \{x_1, x_2, \cdots, x_n\} \tag{2}$$

where x_i represents the text feature vector output by BERT for the i th token, n represents the number of tokens in the text, and X represents the token sequence of the input text. BERT can obtain the corresponding context representation:

$$H = \text{BERT}(X) = \{h_1, h_2, \cdots, h_n\} \tag{3}$$

here, X represents the token sequence of the input text, h_i represents the contextual semantic vector obtained after the i-th token is encoded by BERT, and H represents the complete textual semantic representation.

C. Public Attitude Classification Based on Logistic Regression

After obtaining the BERT text semantic features, this paper employs Logistic Regression to build a classification model for public attitudes toward AI. This method features a simple structure, high computational efficiency, and interpretable classification results, and can serve as a classifier following the BERT semantic features to perform text attitude recognition [5].

For the input text feature vector x, Logistic Regression first computes its linear combination:

$$z = w^T x + b \tag{4}$$

where w represents the model weights, b represents the bias term, and z represents the linear prediction result.

Since this paper classifies public attitudes into three categories—positive, neutral, and negative—multiclass logistic regression is used to calculate the predicted probabilities for each category:

$$P(y = k \mid x) = \frac{\exp(w_k^T x + b_k)}{\sum_{j=1}^{K} \exp(w_j^T x + b_j)} \tag{5}$$

here, $K = 3$ represents the number of attitude categories, k represents a specific category, and $P(y = k \mid x)$ represents the predicted probability that the text belongs to category k. The model ultimately selects the category with the highest predicted probability as the attitude classification result for the text.

Therefore, this paper first uses BERT to extract contextual semantic features from social media text, then employs a logistic regression model to classify public attitudes toward AI, ultimately categorizing the text into three types: positive, neutral, and negative.

D. Evaluation Metrics for the Classification Model

To objectively evaluate the model's performance in identifying public attitudes toward AI, this paper adopts Accuracy, Precision, Recall, and F1-score as the primary evaluation metrics.

Accuracy is used to measure the proportion of samples correctly predicted by the model:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Precision refers to the proportion of text predicted by the model to belong to a certain category that actually belongs to that category:

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

Recall represents the proportion of text that actually belongs to a certain category and is correctly identified by the model:

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

Taking both precision and recall into account, the F1-score is defined as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

For a three-class classification task, the macro-averages of the metrics for each category can be further calculated to reduce the impact of differences in sample sizes across categories on the evaluation results.

In addition, this paper uses a confusion matrix to further examine the classification of positive, neutral, and negative texts, providing a basis for subsequent misclassification analysis.

III. Experimental Analysis of Public Attitudes toward AI Based on Social Media Text

A. Experimental Data Characteristics and Public Attitude Distribution

To validate the effectiveness of the combined BERT and Logistic Regression model in identifying public attitudes toward AI, this paper selected AI-related user comments from social media platforms as experimental data. A total of 2476 text entries were collected from the raw data, with content primarily covering topics such as AI applications, intelligent customer service, intelligent voice interaction, automated services, data security, and user experience. After data cleaning—which involved removing duplicate texts, advertisements, comments lacking substantive semantic content, and those that were too short or whose meaning could not be determined—a total of 356 entries were excluded, resulting in 2120 valid texts, with a valid data retention rate of 85.62%.

During the manual annotation phase, valid texts were categorized into three types—positive, neutral, and negative—based on the primary attitude expressed in the text.

Ultimately, 683 positive texts were identified, accounting for 32.22%; 607 neutral texts, accounting for 28.63%; and 830 negative texts, accounting for 39.15%, as shown in Table 1. It can be observed that the proportion of negative texts is slightly higher than that of positive and neutral texts, which is consistent with the tendency of social media users to express opinions on specific issues. At the same time, there was no extreme sample imbalance among the three categories of texts, meeting the basic requirements for training subsequent classification models.

Table 1. Distribution of public attitudes toward AI by category

Attitude Category	Number of Texts	Percentage
Positive	683	32.22%
Neutral	607	28.63%
Negative	830	39.15%

To avoid excessive influence of category distribution differences on model training, this paper employs stratified random sampling to partition the training and test sets, with approximately 80% of the text used for model training and the remainder for model testing. The final training set contains 1696 text entries, and the test set contains 424 text entries, while maintaining consistency in the proportion of attitude categories across both datasets as much as possible. The specific distributions are shown in Table 2.

Table 2. Category distribution of the training and test sets

Dataset	Positive	Neutral	Negative	Total
Training set	546	486	664	1696
Test set	137	121	166	424
Total	683	607	830	2120

From the above data, the experimental dataset does not show a completely balanced tri-class distribution; there are many more negative texts and few neutral texts. This situation well reflects the imbalance in attitude expressions in social media texts and also evaluates the classification ability of the model under a non-completely balanced data condition.

B. BERT Text Semantic Representations and Attitude Classification Results

To evaluate the efficiency of integrating BERT's contextual semantic features with a Logistic Regression classifier, this study also established four approaches for comparison: TF-IDF+Logistic Regression, TF-IDF+SVM, BERT+SVM and BERT+Logistic Regression. In these methods, TF-IDF is used to represent traditional term frequency statistical features, BERT is used to extract text features containing contextual semantic information, and Logistic Regression and SVM are used as classifiers to perform the three-class classification task, respectively. The experimental results are shown in Table 3.

Table 3. Overall experimental results for different classification algorithms

Classification Method	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
TF-IDF + Logistic Regression	78.77%	78.54%	77.96%	78.21%
TF-IDF + SVM	81.84%	81.66%	81.13%	81.37%
BERT + SVM	85.14%	85.03%	84.68%	84.84%
BERT + Logistic Regression	86.56%	86.49%	86.19%	86.32%

As shown in Table 3, the four models show clear differences in classification performance. TF-IDF+Logistic Regression achieved 78.77% Accuracy and 78.21% Macro-F1, while TF-IDF+SVM improved these values to 81.84% and 81.37%. With BERT-based contextual features, BERT+SVM reached 85.14% Accuracy and 84.84% Macro-F1. BERT+Logistic Regression achieved the best overall performance, with 86.56% Accuracy and 86.32% Macro-F1. Compared with TF-IDF+Logistic Regression, its Accuracy and Macro-F1 increased by 7.79 and 8.11 percentage points, respectively. These results demonstrate that BERT provides more effective semantic representations for social media text classification.

Further analysis of the classification results from the BERT+Logistic Regression model shows that the model correctly identified 367 out of 424 test texts, achieving an overall accuracy of 86.56%. The macro-precision was 86.49%, the macro-recall was 86.19%, and the macro-F1 score was 86.32%, indicating that the model exhibits relatively stable classification performance across the three attitude categories, as shown in Table 4.

Table 4. Classification results of the BERT+Logistic regression model across different attitude categories

Attitude Category	Number of Test Samples	Precision	Recall	F1-score
Positive	137	90.23%	87.59%	88.89%
Neutral	121	83.19%	81.82%	82.50%
Negative	166	86.05%	89.16%	87.57%
Macro Average	424	86.49%	86.19%	86.32%

As shown in Table 4, the F1-score for the positive category reached 88.89%, with a precision of 90.23%, indicating that most of the text classified by the model as positive does indeed exhibit a clear positive bias. The recall for the negative category reached 89.16%, indicating that the model was able to identify most of the text that actually expressed a negative attitude. In contrast, the recall and F1-score for the neutral category were 81.82% and 82.50%, lower than those for the positive and negative categories.

C. Classification Results and Confusion Matrix Analysis

To further analyze the performance of the BERT+Logistic Regression model in distinguishing between different attitude categories, this paper constructed a confusion matrix based on the prediction results of the 424 test texts, as shown in Table 5.

As shown in the confusion matrix, the model correctly classified 367 text samples and misclassified 57. In the positive category, 120 samples were correctly classified, resulting in a correct classification rate of 87.59%; in the neutral category, 99 samples were correctly classified, resulting in a correct classification rate of 81.82%; and in the negative category, 148 samples were correctly classified,

resulting in a correct classification rate of 89.16%.

Table 5. Classification confusion matrix for the test set

Actual Category \ Predicted Category	Positive	Neutral	Negative	Total
Positive	120	8	9	137
Neutral	7	99	15	121
Negative	6	12	148	166
Total	133	119	172	424

D. Discussion

A detailed investigation of the experimental outcomes indicates that various combinations of text features and classification methods result in considerable variations in performance. The accuracy of TF-IDF combined with Logistic Regression is 78.77%, whereas TF-IDF together with SVM increases it to 81.84%. This suggests that the traditional term frequency features can accomplish basic text classification; however, they have some restrictions in comprehending the contextual semantic information. After incorporating the BERT semantic features, the Accuracy of BERT combined with SVM reaches 85.14%, which confirms the beneficial effect of the contextual semantic representation on the text classification of social media. Finally, BERT combined with Logistic Regression obtains an Accuracy of 86.56% and a Macro-F1 of 86.32%, showing a relatively stable general classification ability. For specific categories, the F1-scores of positive and negative texts are both 88.89% and 87.57% respectively, while that of the neutral text is only 82.50%. From the confusion matrix, it can be seen that the misclassifications are mainly distributed between neutral and negative, as well as between neutral and positive, indicating that the texts with indistinct attitudes or having multiple tendencies are more likely to cause classification errors.

IV. Conclusion

This paper conducts classification experiments focused on identifying public attitudes toward artificial intelligence in social media texts. By comparing traditional text features with BERT semantic features, we analyze the classification performance of different models across positive, neutral, and negative text categories. The experimental results show that BERT-based models outperform traditional TF-IDF-based models overall. Specifically, the BERT+Logistic Regression model achieved an accuracy of 86.56% on the test set,and a Macro-F1 score of 86.32%, indicating that contextual semantic representations effectively support the identification of public attitudes. Further analysis of the confusion matrix reveals that the model performs well in distinguishing positive and negative texts, while neutral texts are more prone to confusion with other categories. Therefore, combining BERT with traditional classifiers provides an effective technical approach for analyzing public attitudes toward AI.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] T. K. Balaji, C. S. R. Annavarapu, and A. Bablani, "Machine learning algorithms for social media analysis: A survey," *Computer Science Review*, vol. 40, 100395, 2021.

[2] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: A tertiary study," *Artificial Intelligence Review*, vol. 54, pp. 4997–5053, 2021.

[3] K. Jia *et al.*, "Sentiment classification of microblogs: A framework based on BERT and CNN with an attention mechanism," *Computers & Electrical Engineering*, vol. 101, 108032, 2022.

[4] M. L. Jamil, S. Pais, J. Cordeiro, and G. Dias, "Detection of extreme sentiments on social networks with BERT," *Social Network Analysis and Mining*, vol. 12, 55, 2022.

[5] A. Bello, S. C. Ng, and M. F. Leung, "A BERT Framework for Sentiment Analysis of Tweets," *Sensors*, vol. 23, no. 1, 506, 2023.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).