Linguistic Differences between AI and Human Comments in Weibo: Detect AI-Generated Text through Stylometric Features

Ziqi Li Qi Zhang[†]
School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China
liziqi@uibe.edu.cn zhangqi@uibe.edu.cn

Abstract

LLM-enhanced social robots (LLM-Bots) generate responses similar to human interactions and pose risks to social media platforms. Distinguishing AI-generated texts (AIGTs) from human-written content is important for mitigating these threats. However, current AIGT detection technologies face limitations in social media contexts, including inadequate performance on short texts, poor interpretability, and a reliance on synthetic datasets. To address these challenges, this study first constructs a social media dataset composed of 463,382 Weibo comments to capture real-world interactions between LLM-Bots and human users. Second, a stylometric feature set tailored to Chinese social media is developed. We conduct a comparative analysis of these features to reveal linguistic differences between human-written and AI-generated comments. Third, we propose a lightweight stylometric feature-based self-attention classifier (SFSC). This model achieves a strong F1-score of 91.8% for detecting AI-generated short comments in Chinese while maintaining low computational overhead. Additionally, we provide interpretable criteria for the SFSC in AIGT detection through feature importance analysis. This study advances detection for AI-generated short texts in Chinese social media.

Keywords: AI-generated Text Detection , Natural Language Processing , Stylometric Feature , Deep Learning , Online Social Network

1 Introduction

Recent advances in large language models (LLMs), such as ChatGPT and DeepSeek, have significantly enhanced the capabilities of social bots. These LLM-enhanced social bots (hereafter abbreviated as LLM-Bots) can analyze complex human communication patterns and generate responses indistinguishable from those of real users (Ferrara, 2023). While these human-like interactions bring advantages in improving task efficiency and user interaction, the ensuing threats on social media platforms raise critical concerns. LLM-Bots exacerbate the spread of toxic content, including hostile replies, discrimination, and violent content (Li et al., 2023). Furthermore, LLM-Bots reduce the cost and complexity of generating large-scale disinformation (Hu, 2024). Therefore, they are used to manipulate public opinion through large-scale deployment and high interaction frequency (Wei, 2024).

Accurate detection of social bots is important for maintaining the integrity of social media platforms (Ferrara, 2023). However, LLM-driven manipulation strategies reduce the performance of traditional social bot detection methods by posting on diverse topics, engaging in organic interactions via retweets and replies, and constructing synthetic identities using stolen profiles (Hu, 2024; Yang and Menczer, 2024). Therefore, distinguishing human-written from AI-generated texts (AIGTs) is crucial for LLM-Bot detection.

Currently, AIGT detection technologies have been developed using methods such as watermarking, statistical and stylistic features, pre-trained models, and LLMs as detectors (Fraser et al., 2025; Wu et

[†]Corresponding Author
©2025 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

al., 2025). Despite these advancements, critical challenges persist in AIGT detection on social media platforms. First, existing research focuses on long texts like articles, whereas the detection models exhibit performance limitations when applied to short texts (Mireshghallah et al., 2024; Ma and Wang, 2024), which are prevalent in social media contexts. Second, pre-trained classifiers require substantial computational resources and exhibit poor interpretability. These limitations hinder their practical deployment in dynamic and large-scale environments on social media platforms (Ferrara, 2023). Third, existing research relies on synthetic datasets, lacking real data regarding the interaction dynamics between LLM-Bots and human users on social media contexts (Deng et al., 2024). Moreover, comparative analyses of human-bot linguistic patterns on Chinese social media remain underexplored due to the dominance of English-centric datasets (Fraser et al., 2025).

This study proposes three primary research components in response to these challenges. First, we construct a Chinese-language microblog dataset containing 463,382 comments from Weibo, which captures real-world interactions between LLM-Bots and human users. Second, we develop a stylometric feature set tailored to Chinese social media. Using these features, we conduct a comparative analysis to identify linguistic differences between human-written and AI-generated comments. Third, we propose a lightweight stylometric feature-based self-attention classifier (SFSC) that achieves an F1-score of 91.8% in AI-generated short-comment detection with low computational overhead. We also provide interpretable criteria for AIGT detection through feature importance analysis.

2 Methodology

2.1 Dataset Construction

The dataset construction process followed three systematic stages.

First, we identify a list of LLM-Bot accounts from Sina Weibo's official disclosure. Using the names of these LLM-Bot accounts as search keywords, we collect microblog posts from January to December 2024 that potentially contain human-bot interactions. Initial preprocessing removes duplicate entries and irrelevant content, followed by length filtering that excludes posts containing fewer than 10 or exceeding 500 characters. This process results in 404,581 candidate posts. Subsequently, we exclude suspicious accounts, such as those exhibiting excessively high-frequency bot interactions, retaining 212,905 validated candidate posts. We extract comments for each retained post. To eliminate noise, we remove duplicates and non-substantive comments (e.g., *Repost*). This process results in a post-comment dataset comprising 417,361 human-written comments and 46,021 AI-generated comments.

Second, we apply placeholder substitution to sensitive information. Specifically, user mentions, URLs, and images are replaced with <user>, <url>, and <picture>, respectively. This step preserves the structural integrity of comments for subsequent analysis without compromising user privacy (Sallah et al., 2024).

Finally, we construct a balanced subset of the dataset using stratified sampling to enable comparative analysis between human-written and AI-generated comments. Notably, the dataset excludes replies to comments to prevent human reactions to Known LLM-bots from affecting detection performance. AI-generated comments are categorized into two types: (1) triggered comments, where the original posts mention LLM-Bots, and (2) voluntary comments, where LLM-Bots autonomously reply to posts unrelated to them. To address class imbalance, we randomly sample equal numbers of both AI-generated comment types (18,337 triggered and 18,337 voluntary) and pair them with 33,674 human-written comments. This balanced dataset supports statistical comparisons and model training.

2.2 Feature Set

The stylometric features are designed to identify diverse stylistic patterns in a given text. Building on previous stylometric studies (Opara, 2024; Kumarage et al., 2023; Mindner et al., 2023; Chong et al., 2023) and the linguistic characteristics of the Chinese language, we construct a feature set to distinguish human-written and AI-generated comments on social media platforms. These features are categorized into five dimensions as outlined in Table 1.

CCL 2025

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Category	Feature	Description
Lexical Structure	character	The total number of characters in the text.
	ch_character	The total number of Chinese characters in the text.
	word	The total number of words in the text.
	sentence	The total number of sentences in the text.
	word_sentence_mean	Average number of words per sentence.
	word_sentence_std	Std. deviation of word count per sentence.
	ch_char_word_mean	Average number of Chinese characters per word.
	ch_char_word_std	Std. deviation of Chinese characters count per word.
	ch_char_sentence_mean	Average number of Chinese characters per sentence.
	ch_char_sentence_std	Std. deviation of Chinese characters count per sentence.
Pragmatic Style	stopword	The number of stop words used in the text.
	negation_word	The ratio of negation words to total words.
	assertive_word	The ratio of assertive words to total words.
	degree_adv	The ratio of degree adverbs to total words.
	toned_word	The ratio of toned words to total words.
Readability	error	The ratio of errors to text length.
	level2_char	The ratio of rare characters to total Chinese characters.
	transitional_word	The ratio of transitional words to total words.
	adv_conj	The ratio of adverbs and conjunctions to text length.
Symbol Usage	space	The ratio of spaces to text length.
	total_punct	The ratio of punctuations to text length.
	common_punct	The ratio of frequently used punctuations (， ！ 。？ ：) to text length.
	uncommon_punct	The ratio of punctuations not included in the common punctuation set to text length.
	number	The ratio of numbers to text length.
	emoji	The ratio of spaces to text length.
	placeholders	The ratio of placeholders (<user>, <url>, <picture>) to total words.
Sentiment	sentiment_score	The sentiment score of the text.
	happiness	The number of happiness-related words in the text.
	affection	The number of affection-related words in the text.
	anger	The number of anger-related words in the text.
	sadness	The number of sadness-related words in the text.
	fear	The number of fear-related words in the text.
	disgust	The number of disgust-related words in the text.
	surprise	The number of surprise-related words in the text.

Table 1: Stylometric Features across Five Categories.

First, the lexical structure is examined. Beyond conventional metrics such as character, word, and sentence, we introduce Chinese character-level metrics. These statistical measures include absolute counts, central tendency, and dispersion.

Second, the pragmatic style examines lexical preferences between humans and LLM-Bots. We measure this through frequency analysis of discourse markers, including stop words, negation words, degree adverbs, and the toned words at the end of sentences. We further include counts of assertive verbs (e.g., *think*, *believe*) to quantify the differences in opinion expression between the groups.

Third, we evaluate text readability. Recognizing that most AIGTs rely on high-frequency words (Fraser et al., 2025), we quantify the usage of low-frequency Chinese characters. To address cases where LLMs employ decoding strategies that prioritize lexical diversity, we assess coherence in syntac-

Proceedings of the 24th China National Conference on Computational Linguistics, pages 842-851, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

tic connectors, including adverbs, conjunctions, and transitional words. Additionally, using the *pycorrect* library in Python, we quantify spelling and grammatical errors. AIGT typically shows higher accuracy compared to human-written texts (Fraser et al., 2025; Mindner et al., 2023).

Fourth, we quantify the symbol usage in social media short texts. In addition to standard punctuation marks, our analysis includes counts of whitespace, numerical figures, emojis, and placeholders (<user>, <url>, and <picture>).

Sentiment is the fifth category. Given the observed tendencies of AIGT toward restrained negative sentiment and class imbalance in emotional word usage (Opara, 2024), we employ the *cncenti* library to calculate both overall sentiment scores and the frequencies of seven predefined emotion categories.

2.3 Stylometric Feature-based Self-attention Classifier

We propose a stylometric feature-based self-attention classifier (SFSC) to distinguish human-written from AI-generated social media comments. Figure 1 illustrates the proposed SFSC architecture.

First, we apply Z-score normalization to the 34 stylometric features. This step ensures uniform scaling across features while preserving the original distribution. Following standardization, the SFSC incorporates a self-attention mechanism to dynamically weight the importance of individual features based on their contextual relationships. This attention mechanism enables the model to identify subtle interactions between features that reflect linguistic and structural patterns unique to each type of comment. Subsequently, the weighted features are input into a multilayer perceptron (MLP) composed of two dense layers: a 256-node layer followed by a 128-node layer with ReLU activation for nonlinear transformation. The final layer employs a binary cross-entropy loss function to classify unidentified comments. To prevent overfitting, dropout regularization (rate = 0.1) is applied to dense layers.

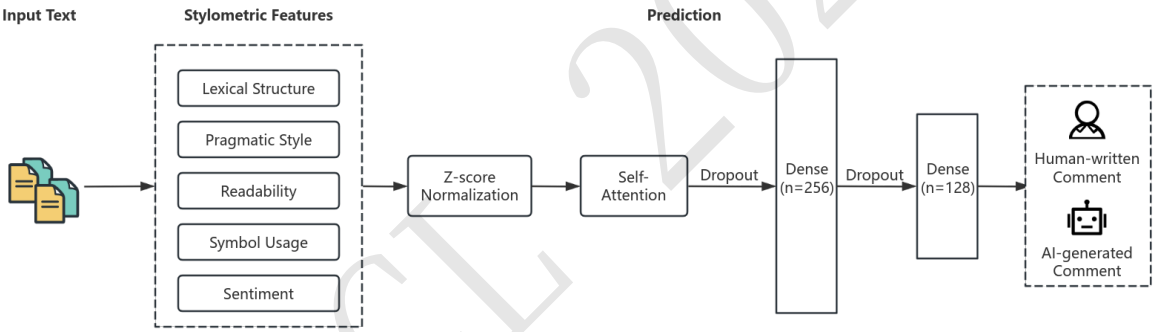


Figure 1: Architecture of the SFSC.

3 Experiments

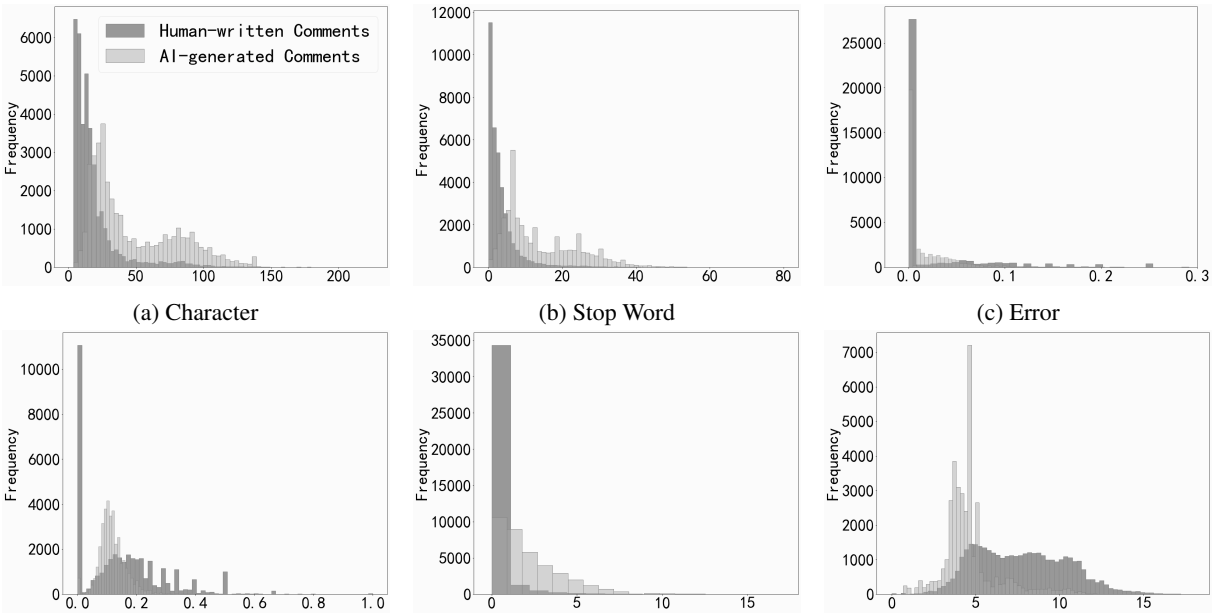
3.1 Comparison Results

To visualize the stylometric differences between human-written and AI-generated social media comments, Figure 2 depicts histograms comparing the frequency distributions of selected features from each category.

In Figure 2 (a), the distribution of character counts shows that AI-generated comments are generally longer than human-written comments. However, this still aligns with the short-text nature of social media, as over 70% of comments are within 60 characters. Figure 2 (b) demonstrates a preference for stop words in AI-generated comments, indicating systematic adherence to syntactic norms. The distribution patterns in Figure 2 (c) reveal significant differences in textual error rates. AI-generated text exhibits low error rates for misspelling and grammatical mistakes, with lower variance in error distribution. Conversely, human-written comments display a bimodal distribution. Most texts are error-free, while a distinct subset exhibits comparatively higher error frequencies. This bimodal distribution may reflect the intentional use of non-standard language, such as internet slang and informal expressions. Regarding

CCL 2025

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China



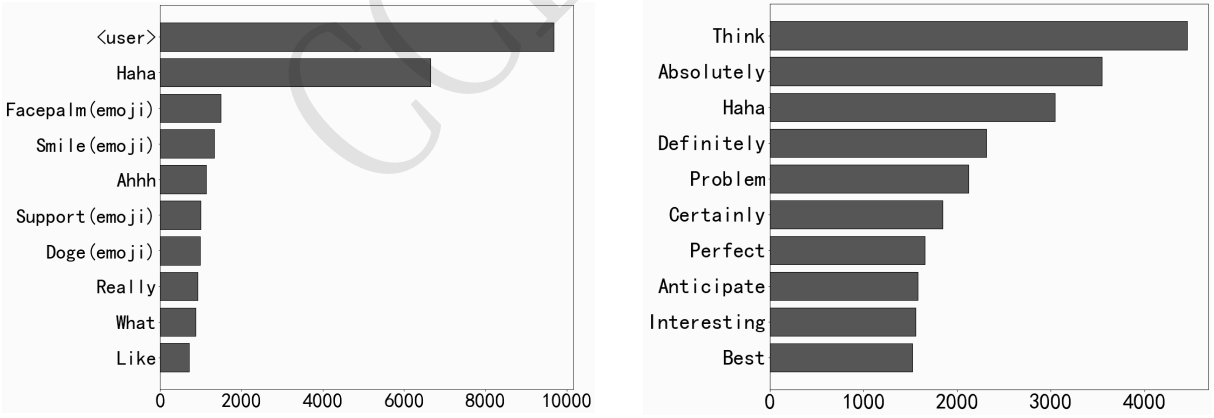
(a) Character (b) Stop Word (c) Error

(d) Punctuation (e) Affection (f) Response Rate

Figure 2: Comparative histograms of human-written and AI-generated comments across stylometric features and response rate.

punctuation usage in Figure 2 (d), human-written comments exhibit three distinct clusters: no punctuation, normative usage, and excessive punctuation application. By contrast, AIGTs approximate a normal distribution in punctuation usage frequency, adhering more closely to syntactic norms. The analysis of emotional expression in Figure 2 (e) focuses on affection-related lexicon, such as *love* and *appreciate*. LLM-Bots demonstrate excessive use of this emotional category compared to human frequencies.

Besides stylometric features, Figure 2 (f) shows differences in response rate, measured as the logarithm of time intervals between post publication and comment submission. While LLM-Bots generally respond faster than humans, some are adopting delayed response strategies to mimic human behavior. These findings highlight the importance of stylometric features in the effective detection of LLM-Bot accounts.



(a) Human-written Comments (b) AI-generated Comments

Figure 3: Top 10 High-Frequency Words (Translated from Chinese).

Lexical preference analysis presented in Figure 3 compares the top 10 high-frequency words between human-written and AI-generated comments. Human-written comments primarily include interaction-oriented elements such as user mentions (@<username>) and emojis. These elements suggest that humans tend to use platform-specific communication. In contrast, AI-generated texts exhibit semantic

Model	Accuracy	Precision	Recall	F1-Score	Training Time(s)	Testing Time(s)
LR	0.708	0.788	0.570	0.661	<u>1.629</u>	**0.001**
KNN	0.581	0.545	**0.969**	0.698	115.681	11.803
RF	0.569	**0.958**	0.144	0.250	94.004	0.957
NB	0.584	0.563	0.750	0.643	**0.285**	<u>0.002</u>
RoBERTa	**0.931**	0.902	<u>0.968</u>	**0.934**	663.090	8.432
SFC	0.906	0.898	0.917	0.907	23.044	4.711
SFSC	<u>0.917</u>	<u>0.911</u>	0.925	<u>0.918</u>	31.240	4.926

Table 2: Performance Comparison on AIGT Detection.

uniformity through frequent use of verbs *think* and absolute adverbs including *absolutely*, *definitely*, and *certainly*.

This comparison between human-written and AI-generated comments provides behavioral cues for human observers aiming to identify AIGT.

3.2 Classification Results

Initially, we compare the performance of our proposed SFSC with four machine learning classifiers, i.e., Logistic Regression (LR), KNN, Random Forest (RF), and Naive Bayes (NB), as well as the pre-trained language model RoBERTa. We employ an 8:2 split of the samples into the train set and test set while preserving a balanced distribution of human-written and AI-generated comments in both sets. In the training phase, our SFSC and the RoBERTa employ the Adam optimizer with an initial learning rate of 1×10^{-4}. Training proceeds for 20 epochs with a batch size of 32. For the machine learning models, we use TF-IDF to extract features and perform a 5-fold cross-validation method on the training set to optimize the hyperparameters. Performance evaluation incorporates metrics such as accuracy, precision, recall, the F1-score, training time, and testing time.

Table 2 displays the comparison results between our SFSC and the baseline models, with optimal and sub-optimal performance highlighted in bold and underlined, respectively. Our SFSC demonstrates competitive detection performance, with results in accuracy and F1-score second only to the state-of-the-art RoBERTa. Notably, our SFSC exhibits high efficiency, with both training time and testing time substantially lower than those of RoBERTa. This efficiency gap makes SFSC more suitable for the dynamic social media platforms where rapid processing is essential. Furthermore, compared to the black-box pre-trained model, SFSC likely possesses greater potential for interpretability. This interpretability facilitates the understanding of the model's decision-making process, further enhancing its practical deployment. Therefore, SFSC achieves a well-balanced performance in detection capability, efficiency, and interpretability, highlighting its advantages in real-world social media AIGT detection.

Additionally, the penultimate row in Table 2 presents the performance of the ablation model, i.e., the stylometric feature-based classifier (SFC) without the self-attention mechanism. The results reveal that compared to SFC, the SFSC achieves marked improvements in accuracy, precision, recall, and F1-score performance, despite a marginal reduction in computational efficiency.

Besides, a detailed error analysis identifies the difference in detection rates between triggered and voluntary AI-generated comments. Specifically, 5.26% triggered comments are misclassified as human-written, while voluntary comments exhibit a higher false-negative rate of 11.94%. To investigate this performance discrepancy, we compare the triggered, voluntary, and human-written comments in Table 3. We observe that triggered comments, typically constrained by specific patterns such as question-and-answer interactions, often exhibit structural similarities. This consistency makes them more easily identifiable by our SFSC. In contrast, voluntary comments appear to have greater naturalness and align more closely with human-written comments in stylometric features. This similarity poses significant classification challenges, resulting in a higher false negative rate.

To evaluate classification performance across different LLM-Bots types, we examine four subcategories with test data exceeding 100 instances. The descriptions, account counts, and comment counts

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	Triggered Comment Context	Voluntary Comment Context
Weibo	我跟哪位微博好友最匹配<user> (Which of my Weibo friends am I most compatible with <user>)	常常在思考为什么人与人之间的羁绊一下子就结束了 (I often find myself pondering: why do the bonds between people end so abruptly)
Comments Pending Detection	**Triggered Comment 1**: 看样子，你和<user>最配！就像北极冰山的融化，彼此带来温暖和甜蜜。赶紧和他展开更多交流吧！ (Looks like you and <user>are the perfect match! Just as polar ice thaws, you bring each other warmth and sweetness. Start building that connection now!)	**Human-written Comment**: 其实我比任何人都希望回到从前的某段时光。我只是一直疯狂压抑自己的感情，不断重复着对自己说：一直向前，永远向前。 (Truth be told, I yearn to revisit certain moments from the past more deeply than anyone. I've just been frantically holding back these emotions, relentlessly repeating to myself: 'Keep moving ahead—now and always.)
	Triggered Comment 2: 看起来你和<user>这个宝宝最配哦！他们的可爱和你的幽默一定能擦出火花！快去找<user>互动吧！ (It seems you and <user>are the perfect match! Your humor and their charm are bound to spark exciting chemistry. Go ahead and strike up a conversation with <user>!)	**Voluntary Comment**: 因为人生就像乘公交，有人上车，就有人下车，到站了，自然就散了 (Because life unfolds like riding a bus—passengers come and go, and bonds dissolve as naturally as reaching each stop.)

Table 3: Examples of Triggered, Voluntary, and Human-written Comments.

Category	Accounts	Comments	Description
Weibo Reply Bot	2	3329	Official LLM-Bots of Weibo.
Functional Interactive Bot	4	1789	LLM-Bots that provide automated services such as psychological assessments and account evaluations.
Fictional Character Imitator	90	1123	LLM-Bots that imitate the mannerisms of fictional characters from TV shows, games, and other media.
Blogger Assistant	115	117	LLM-Bots with personalized reply styles that respond to comments on behalf of the bloggers.

Table 4: Summary of LLM-Bot Categories and Their Counts.

of these subcategories are presented in Table 4. Figure 4 illustrates their recall values. Functional Interactive Bots achieve the highest recall, likely attributed to their task-oriented style. Conversely, Blogger Assistants demonstrate the lowest recall. These LLM-Bots exhibit heterogeneous stylometric features resulting from personalized configurations and diverse knowledge bases.

To evaluate SFSC's adaptability to evolving AI generation techniques, we conduct monthly performance evaluations using data comprising over 50 samples per month. As shown in Figure 5, early-stage data exhibit lower accuracy and F1-scores, attributed to limited training samples capturing incipient AIGT style. Despite the deployment expansion of upgraded LLM-Bots by multiple institutions, SFSC has maintained stable performance since July. This suggests that stylometric features can provide robust detection capabilities in real-world applications. However, the need for dynamic feature libraries to accommodate evolving LLMs remains critical.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 842-851, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

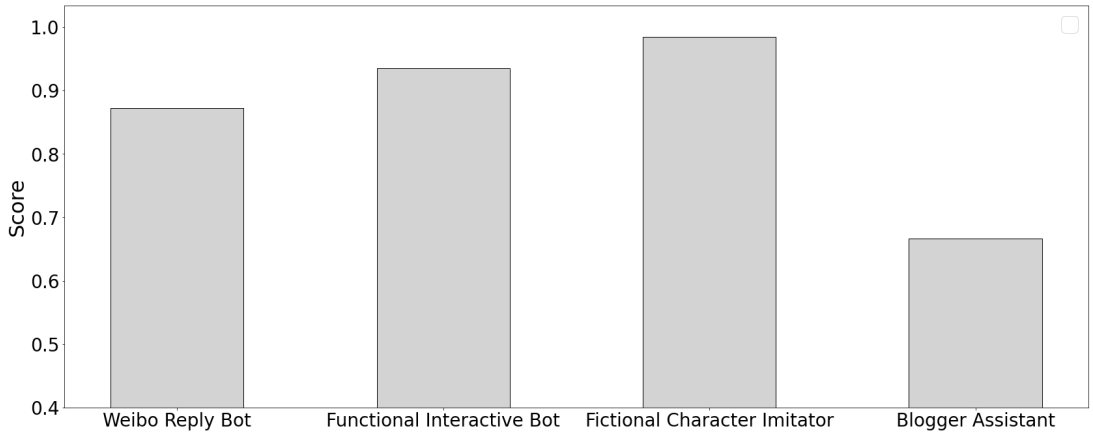


Figure 4: Recall Rates across Four LLM-Bots Types.

CCL 2025

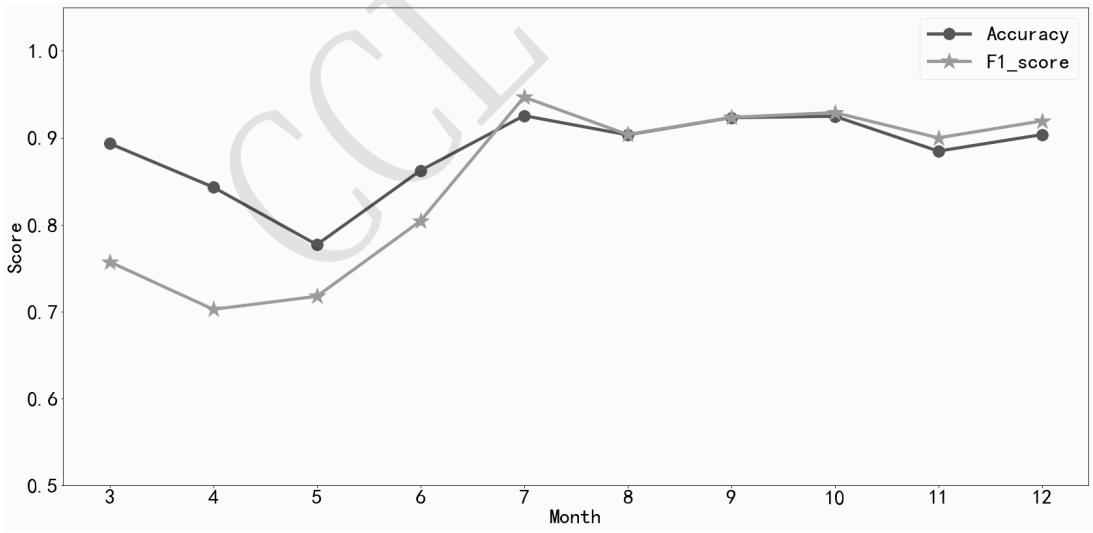


Figure 5: The Monthly Results of Accuracy and F1-score.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 842-851, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3.3 Feature Importance

We quantify relative feature importance by calculating the average attention weights of the self-attention mechanism. The ten most influential features are shown in Figure 6.

Sentiment-related features, especially those quantifying emotion of affection, happiness, and surprise, serve as the primary discriminators in AI-generated comments detection. Lexical structure constitutes the secondary influential category. The distributions of sentences, words, characters, and Chinese characters demonstrate high weights in classification. The preference for stop words is also crucial to distinguish human-written comments from AI-generated ones. Notably, features related to symbol usage rank comparatively lower in importance. This finding contrasts with a previous English-language study where punctuation serves as a key marker (Kumarage et al., 2023). This discrepancy may stem from differences in punctuation conventions between Chinese and Western social media contexts.

The importance analysis emphasizes the necessity of language-specific features for localized AIGT detection. The analysis also improves model interpretability for human observers and platform managers.

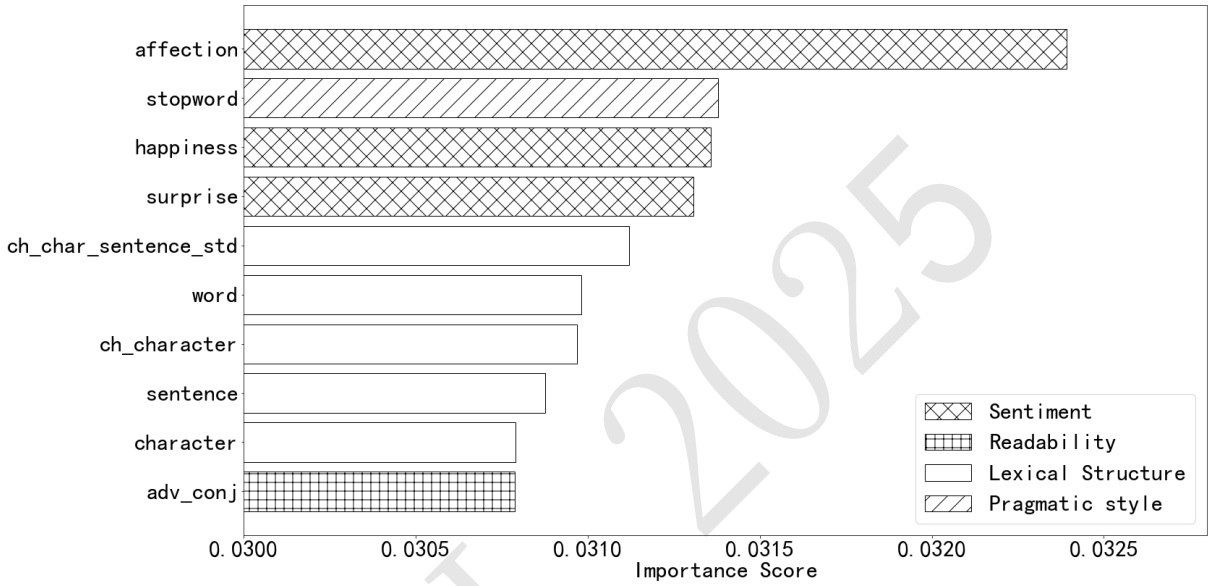


Figure 6: Feature Importance.

4 Conclusion

This study constructs a dataset comprising posts and comments from interactions between humans and LLM-Bots on the social media platform Weibo. By establishing 34 stylometric features across five categories, we identify statistical differences between human-written and AI-generated short texts in Chinese social media contexts. Based on these empirical findings, we propose a stylometric feature-based self-attention classifier to address the challenge of detecting AIGT in short social media comments. The SFSC model achieves a strong F1-score of 91.8% with low computational overhead, demonstrating its effectiveness and efficiency in distinguishing stylistic differences in short-text scenarios. The subsequent feature importance analysis reveals critical features in the AIGT task. This analysis improves model interpretability for SFSC and offers actionable insights for refining detection strategies.

The current dataset is confined to Weibo data due to the availability of verified LLM-Bot accounts. Future research directions include cross-platform validation in platforms like Xiaohongshu, where AIGTs are increasing rapidly.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62106047) and Scientific Research Laboratory of AI Technology and Applications, University of International Business

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

and Economics.

References

Alicia Tsui Ying Chong, Hui Na Chua, Muhammed Basheer Jasser, and Richard T.K. Wong. 2023. Bot or human? Detection of deepfake text with semantic, emoji, sentiment and linguistic features. In *2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)*, pages 205–210.

Shenli Deng, Fan Wang, and Haowei Wang. 2024. Identification methods of artificial intelligence generated content in online communities. *Documentation,Information & Knowledge*, 41(02):28–38+149.

Emilio Ferrara. 2023. Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*, 28(6), Jun.

Kathleen C Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2025. Detecting ai-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, 82:2233–2278.

Yong Hu. 2024. Ai-driven disinformation: Present and future. *Nanjing Journal of Social Sciences*, (01):96–109.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines.

Siyu Li, Jin Yang, and Kui Zhao. 2023. Are you in a masquerade? Exploring the behavior and impact of large language model driven social bots in online social networks.

Shixuan Ma and Quan Wang. 2024. Zero-shot detection of LLM-generated text using token cohesiveness. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17538–17553, Miami, Florida, USA, November. Association for Computational Linguistics.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human- and ai-generated texts: Investigating features for ChatGPT. In Tim Schlippe, Eric C. K. Cheng, and Tianchong Wang, editors, *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, pages 152–170, Singapore. Springer Nature Singapore.

Niloofar Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2024. Smaller language models are better black-box machine-generated text detectors.

Chidimma Opara. 2024. Styloai: Distinguishing ai-generated content with stylometric analysis. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, editors, *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 105–114, Cham. Springer Nature Switzerland.

Amine Sallah, El Arbi Abdellaoui Alaoui, Said Agoujil, Mudasir Ahmad Wani, Mohamed Hammad, Yassine Maleh, and Ahmed A. Abd El-Latif. 2024. Fine-tuned understanding: Enhancing social bot detection with transformer-based classification. *IEEE Access*, 12:118250–118269.

Junbin Wei. 2024. The construction of the pdrr system for intelligent generation and risk prevention of online public opinion. *Nanjing Journal of Social Sciences*, (06):76–87.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338, 03.

Kaicheng Yang and Filippo Menczer. 2024. Anatomy of an ai-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, 4, May.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China