Journal of King Saud University - Computer and Information Sciences 36 (2024) 102010

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Full Length Article

Punctuation and lexicon aid representation: A hybrid model for short text sentiment analysis on social media platform

Zhenyu Li [a], Zongfeng Zou [b,*]

[a] School of Cultural Heritage and Information Management, Shanghai University, Shanghai, China
[b] School of Management, Shanghai University, Shanghai, China

ARTICLE INFO

Keywords:
Sentiment analysis
Short text classification
BERT representation
Attention mechanism
Social media mining

ABSTRACT

Sentiment analysis measures user experience on social media. With the emergence of pre-trained models, text classification tasks have become homogeneous, without a significant improvement in accuracy. Therefore, we present a hybrid model called PLASA to classify the sentiment polarity of short comments, particularly barrages. PLASA introduces a collaborative attention module that integrates information about relative position and knowledge from summarized lexicons to better adjust the relationship between word representations. Our model is evaluated using three new curated sentiment analysis datasets: SentiTikTok-2023 (4613 items), SentiBilibili-2023 (7755 items), and SentiWeibo-2023 (5614 items). Although the comment length varies across datasets, all maintain a consistent punctuation percentage at approximately 13%. Consequently, PLASA with the optimal combination demonstrates notable performance improvements compared to both the baseline and commonly used models. It achieves micro-F1 scores of 93.94%, 90.34%, and 88.79% on the respective datasets. We also observed that the representation capacity of the pre-trained model decreases as the text length increases. Moreover, the proposed collaborative attention module effectively addresses this limitation, as confirmed by our ablation study.

1. Introduction

Recently, short videos and live streams have become integral components of modern-day life. People worldwide frequently post content, exchange information, and complete transactions on social media platforms (Koukaras et al., 2020). At the same time, the progress of online services and the widespread adoption of smartphones have profoundly transformed various industries. For instance, in marketing, live streaming allows for up-close demonstrations and interactive Q&A sessions, overcoming the constraints of time and space. Nonetheless, there is an imbalanced relationship between users and staff, resulting in numerous fans not receiving timely concerns and solutions. Hence, it is necessary to automatically analyze user experience and satisfaction, which can be captured through online comments (Park et al., 2020). Human engagement on social media platforms demands an investment of time, emotional involvement, and sometimes financial resources. Commenting behaviors are displayed by users who are already interested in posts, videos, and live streams. However, not all interest-driven behaviors are positive; participants can be divided into two conflicting groups: fans and anti-fans. Although their comments sometimes are neutral, the criterion for distinguishing these groups lies in the sentiment binary polarity of the comments, forming the basis for subsequent tasks, such as election predictions (Tumasjan et al., 2010), policy evaluation (Bai et al., 2022), and depression diagnosis (Stamatis et al., 2022).

Earlier sentiment analysis approaches involved directly matching empirical summarized lexicons to calculate sentiment scores. Additionally, there are many lexicon-based tools available for sentiment analysis, e.g., Linguistic Inquiry and Word Count (LIWC),[1] SentiStrength,[2] Sentiment Analysis and Cognition Engine (SEANCE).[3] However, these existing tools often lack flexibility and show subpar performance when applied to social media data (Bonta et al., 2019). Natural language takes on various forms ranging from formal academic statements to informal social media comments, depending on factors

* Corresponding author.
E-mail addresses: sometimesrains@shu.edu.cn (Z. Li), zfzou@mail.shu.edu.cn (Z. Zou).
[1] https://www.liwc.app.
[2] https://sentistrength.wlv.ac.uk.
[3] https://www.linguisticanalysistools.org/seance.html.

https://doi.org/10.1016/j.jksuci.2024.102010
Received 7 November 2023; Received in revised form 8 March 2024; Accepted 9 March 2024
Available online 11 March 2024

such as context, medium, and timeliness (Balahur and Perea-Ortega, 2015). Enhancing the lexicon-based approach to effectively address various scenarios, such as assigning scores to emotional words (Jurek et al., 2015) and continually updating the lexicon (Alshaabi et al., 2022), is crucial. On the other hand, driven by the advancements in AI, many scholars focused on using machine learning and deep learning for sentiment analysis in the context of social media (Choi et al., 2014; Bai et al., 2022; Qian et al., 2022; Alwehaibi et al., 2022; Govindan and Balakrishnan, 2022). However, there is a lack of consistency in the results obtained by different approaches on the same dataset or by the same approach on different datasets (He, Yin, & Zheng, 2022). Recently, pre-trained models have gained widespread popularity in representing texts (de Andrade et al., 2023), leading to advanced results in sentiment analysis for social media (Melton et al., 2022; Suhaimin et al., 2023). Nevertheless, the high noise in informal language locks down its forward performance.

While previous sentiment analysis approaches have yielded rewarding outcomes, several significant challenges remain to be addressed. Firstly, the majority of studies have been conducted on English-language social media platforms (e.g. Twitter and Reddit). During the Web 2.0 era, the development of Chinese social media has positioned it as a global leader. The use of platforms like TikTok, Bilibili, and Sina Weibo has become an integral component of modern-day life for Chinese people, making them promising sources for sentiment analysis. However, sentiment analysis studies for Chinese social media are scarce. In particular, the utilization of barrage, commonly referred to as 'danmu' (Chinese translation: 弹幕) or 'bullet screen', has become prevalent on Chinese video and live-streaming platforms in recent years (Chen et al., 2019; Zhang and Cassany, 2020; Chen, 2020). Barrage comments are distinguished by their ultra-short length, frequent occurrence, and rapid scrolling, which present challenges for analysis (Hao et al., 2019). Secondly, the sentiment of a sentence is influenced by the interplay between semantics and punctuation, with a notable phenomenon observed in barrages. For example, in an e-commerce live streaming, the comments "These shoes are nice" and "These shoes are nice?" are opposite in terms of sentiment polarity. Despite both sentences containing the positive term "nice", the presence of the question mark "?" in the second comment accents an astonished and skeptical tone. This leads to reversing the sentiment into a negative. Let's consider another scenario where determining the sentiment of the movie review "Tears" is challenging. This is because tears can be evoked by both sadness and excitement, depending on the movie's plot. However, the presence of "!" in "Tears!" undoubtedly conveys a positive signal, as the exclamation mark signifies an exciting linguistic context even without additional information. However, previous studies have not systematically investigated the emergence of punctuation in social media. Last but not least, the current sentiment analysis methods heavily rely on the text representation of pre-trained model BERT (Devlin et al., 2019), while overlooking the potential of lexicon-based features (Alshaabi et al., 2022). Nevertheless, prior literature highlights that emotional words, negations, and punctuation within sentences significantly influence sentiment analysis (Cureg et al., 2019; Dhaoui et al., 2017).

Motivated by these limitations, we propose a hybrid model called PLASA (Punctuation-location and Lexicon-query Attentions for Sentiment Analysis), which utilizes human perception and cognition as references to improve sentiment analysis. The pre-trained model BERT serves as the foundation for the PLASA. Additionally, it integrates information about the word spacing from punctuation and word matching in widely valid lexicons to regulate the relationships of word representations in sentence sequence. Our proposed approach offers benefits to platforms, technology service companies, and government agencies for detecting user sentiment and intergroup sentiment on social media. It also assists content creators in modifying content to appropriately address negative comments that could be harmful and spreading. Below is a summary of our research's significant contributions:

- For barrages' sentiment analysis, we curate two datasets: SentiTikTok-2023 and SentiBilibili-2023. These datasets are sourced from TikTok and Bilibili, two major social media platforms with limited available datasets in academia and industry. Moreover, we partially address the bias due to automatic labeling in the publicly available dataset Weibo_Senti_100k from Sina Weibo. This effort led to the creation of a new dataset named SentiWeibo-2023.
- This is an innovative study that analyzes the sentiment of comments on social media platforms taking human perception and cognition as references. Our study aims not only to enhance the accuracy of sentiment analysis but also to capture the most authentic information absorption during user interactions with massive data streams.
- We propose a hybrid model called PLASA that combines state-of-the-art NLP techniques while preserving valuable insights from traditional sentiment analysis methods as optimization guidance. PLASA integrates punctuation-location attention and lexicon-query attention mechanisms to adjust the relationship between word representations in sequence. A classifier is then utilized to learn features, resulting in outstanding task performance.
- Our model PLASA outperforms commonly used models, baseline models, and state-of-the-art (SOTA) models on three datasets (SentiTikTok-2023, SentiBilibili-2023, and SentiWeibo-2023). Additionally, the necessity of each module in PLASA is ascertained through ablation experiments. Simultaneously, we observed that the shorter the text, the more effective the punctuation-location attention. Conversely, the longer the text, the more pronounced the lexicon-query attention.

2. Related work

2.1. Lexicon-based sentiment analysis

The initial approach to sentiment analysis was formulated by extracting linguistic cues of keywords from text. LIWC is the commonly used and extensively validated tool for calculating various text features, including a significant feature set linked to human affective processes. The first version of the LIWC program and its lexicon were released in 2001 (Pennebaker et al., 2001), with subsequent major updates to the lexicon in 2007 and 2015. In Section 4.2.2, we employ the Chinese simplified LIWC-15 lexicon. Over the past two decades, numerous studies have explored sentiment analysis by leveraging the LIWC tool or lexicons (Tumasjan et al., 2010; del Pilar Salas-Zárate et al., 2014; Bai et al., 2022). Beyond that, many researchers have been willing to create and share their lexicons to facilitate sentiment analysis (Chiarello et al., 2020; Neviarouskaya et al., 2011), including the development of lexicons with automatic updating features (Alshaabi et al., 2022). The original lexicon-based approach simply quantified sentiment by the counts of positive words minus the counts of negative words, which proved to be unreliable when applied to texts with rich emotional words (Choi et al., 2014). Additionally, assigning different emotional intensity values to each word in the lexicon and optimizing the combination function has been shown to be more effective (Jurek et al., 2015). Nonetheless, methods relying entirely on manual design inherently suffer from limitations in terms of speed and accuracy. Owing to the rapid evolution of machine learning, sentiment analysis in social media later integrated LIWC features and machine-learning classifiers, resulting in general advancement in this field. Furthermore, a lot of research has focused on developing optimal feature sets to improve classification performance. For example, Gaston et al. (2018) utilized the steady-state genetic algorithm (SSGA) to identify the most effective feature set of LIWC. Similarly, Crossley et al. (2017) employed stepwise discriminant function analysis (DFA) to remove irrelevant LIWC features for sentiment classification. While the lexicon-based method, even after feature selection, can achieve satisfactory performance, its accuracy remains below that of the AI-based method.

2.2. AI-based sentiment analysis

For computer programs to comprehend natural language, word embedding or representation is essential (Zuhra and Saleem, 2023). Sentiment analysis, a fundamental task in NLP, has seen the evolution of word embedding techniques. Using word frequency statistics was a plain idea for text vectorization in the early stages. Martineau and Finin (2009) employed the bag-of-words model, using both unigrams and bigrams to represent movie reviews, and emphasized the importance of words based on TFIDF. While these encoding schemes were popular in the past work, they did not consider word context. Compared to the bag-of-words model, Word2Vec captures the semantic relationships through context and represents words as dense vectors. Importantly, the introduction of the concept of word distance contributes to a comprehensive advancement in sentiment analysis tasks (Chakraborty et al., 2020). Moving beyond Word2Vec, another well-known word embedding model is Glove (Pennington et al., 2014). This model constructs a co-occurrence matrix from the corpus to capture not only local word context but also broader semantic associations, further enhancing performance in sentiment polarity classification (Qian et al., 2022; Sharma et al., 2017). As both Word2Vec and Glove are static representations, the generated embedding of each word remains fixed after model training. The transformer architecture was originally introduced to tackle sequence-to-sequence problems (Vaswani et al., 2017). Inspired by transformer architecture, the pre-trained BERT model was proposed to generate dynamic word embeddings. Having been trained on multiple large-scale corpora, the BERT model shows superior ability in capturing word semantics, thereby significantly improving the performance of sentiment analysis (Zhao and Yu, 2021; Li et al., 2021a; Melton et al., 2022). After the success of BERT, several transformer-based models have been developed, incorporating various extensions that have shown remarkable effectiveness across a range of NLP tasks (Ahmed et al., 2022). However, it's important to note that none of these models have specifically focused on enhancing strategies for short text classification.

While achieving high accuracy in text classification requires excellent representation, the choice of classifier is equally crucial. A variety of machine-learning classifiers are used for sentiment analysis, e.g. Random Forests (RF) (Al Amrani et al., 2018), Support Vector Machine (SVM) (Liu et al., 2017), and Extreme Gradient Boosting (XGBoost) (Hama Aziz and Dimililer, 2021). Among these classifiers, each armed with a distinct feature extraction method, widely employed in academia and industry, there is no solitary classifier that uniformly surpasses others across all datasets (Alessia et al., 2015; Chakraborty et al., 2020; Dangi et al., 2022). Compared to machine-learning classifiers, deep-learning classification networks, such as Recurrent Neural Network (RNN), excel in processing sequences of word embeddings. Puh and Bagić Babac (2023) conducted a comparison experiment involving various AI-based sentiment analysis approaches, revealing that the optimal one was a Bidirectional Long Short-Term Memory (BiLSTM) network.

2.3. Hybrid model for sentiment analysis

Lexicon-based approaches always assign constant sentiment polarities and scores to keywords, which ignore the discrepancies of the same words encountered in different contexts. Moreover, the lexicon updates lag behind the dynamic shifts of social media language. In contrast, AI-based sentiment analysis encounters challenges in acquiring universal knowledge, resulting in poor transferability across different datasets. The integration of lexicon-based and AI-based approaches, often referred to as a hybrid model, holds the potential to further enhance the performance of sentiment analysis (Dhaoui et al., 2017; Alessia et al., 2015). In the study by Tanna et al. (2020), varying sentiment valences were assigned to words within different emotional categories (e.g. love, happiness, and excitement). This incorporation of lexicon-based features and their corresponding valences helps the classifier achieve more accurate classification. Similarly, Geng et al. (2020) calculated TFIDF values for each LIWC feature, distinguishing their importance and adjusting their weights within the model. These hybrid models process data by incorporating auxiliary information and moderating internal relationships, but they lack a connection to word embeddings. In Rinaldi et al. (2017), the authors encoded LIWC features as one-hot vectors and merged them with word embedding vectors to analyze the sentiment of dialogues. Even though the results weren't as successful as hoped, it was indeed a courageous attempt for the hybrid model. Instead of simple concatenation, the integration of the lexicon-based and AI-based approaches should focus on exploring the mapping relationship between features and representations. Pathak et al. (2021) introduced a topic attention mechanism for sentence-level sentiment analysis. They multiplied the extracted topic attention vectors with the sentence representation obtained from Word2Vec and Glove, then input to the LSTM network for classification. However, it is difficult to extract diverse topics from short comments on short videos and live streams. Currently, the combination of BERT representation and BiLSTM classification network is a reasonable selection for most NLP tasks. On social media platforms, ultra-short barrages might lack the necessary semantic information for effective sentiment analysis, while longer comments could introduce more noise, posing challenges in determining sentiment polarity. Considering this, hybrid models need to focus their attention on features tailored for sentiment analysis.

3. Research objective

The main objective of this research is to introduce a hybrid model that integrates punctuation information and lexicon knowledge for classifying the sentiment polarity of short comments on social media platforms. It is worth noting that most of the existing sentiment analysis approaches have been conducted on English-language datasets, and ignore the combination of punctuation and lexicons with the pre-trained model. To achieve this, our sub-objectives are as follows: (1) To establish publicly available sentiment analysis datasets for Chinese social media. (2) To explore the role of punctuation and lexicon in expressions of social media comments. (3) To design a collaborative attention module integrating punctuation information and lexicon knowledge within the hybrid model. (4) To validate the effectiveness of our hybrid model and the novel attention mechanisms.

4. Methodology

In this segment, we present the overall framework of our novel PLASA model. This top-down hybrid model incorporates the functional components (illustrated in Fig. 1), including the comment representation module, collaborative attention module, and classification module.

4.1. Comment representation module

To enable the computer to understand semantics and analyze sentiments in comments, the first step involves converting words and sentences into vectors. Word2Vec, fastText, and Glove are all content-based word embedding models that require training on a large self-prepared corpus adapted to the specific domain (Onan, 2021). Because the data streams (contexts) on social media platforms are sometimes repetitive and at other times lack coherence, the performance of these self-trained models is limited. More importantly, these models cannot dynamically adjust their representation according to different contexts after training. By contrast, the pre-trained BERT model, having undergone unsupervised learning on multiple large-scale corpora, is capable of generating semantically rich and contextually sensitive embeddings. Subsequently, fine-tuning the representations based on a specific dataset or operating

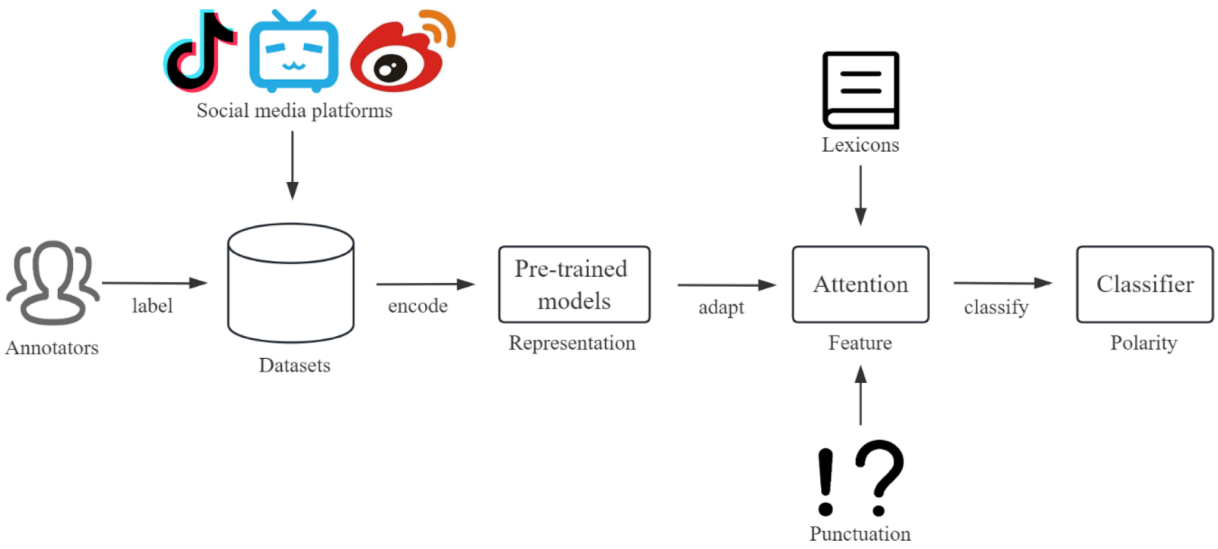

Fig. 1. The illustration of our pipeline for sentiment analysis.

without fine-tuning leads to a relatively high level of performance on downstream tasks. In contrast to traditional Chinese segmentation tools like jieba,[4] the Chinese BERT-base model treats each Chinese character as an individual word (Yu et al., 2022), simplifying the encoding process. Since social media is one of the primary sources for training the Chinese BERT-base model, it is capable of recognizing not only commonly used Chinese characters but also English, Japanese, and numerical characters that frequently appear in social media interactions. Fig. 2 exemplifies the process of generating embeddings using the pre-trained BERT model.

4.1.1. Word and punctuation encoder

Let the length of a comment be denoted as m, and the length required for its encoding as n ($n \geq m + 2$). BERT's tokenizer adds the 'CLS' token before the comment, benefiting downstream classification tasks. Additionally, the 'SEP' token is placed after the comment for text segmentation within the same entry (de Andrade et al., 2023). To achieve a comprehensive representation of comments with varying lengths, the value of n is often set to match the length of the longest comment. Consequently, the majority of comments in a dataset need to use the 'PAD' token for padding. These reasons are why n is at least 2 greater than m. Another relatively rare occurrence is using the 'UNK' token to consistently encode words that cannot have a corresponding match in the encoding dictionary of the Chinese BERT-base model.

In social media, it is common for users to mix Chinese and English punctuation. Although they may be nearly identical in terms of semantics and typically used unconsciously, their string forms cannot be considered equivalent in computer programs. Hence, we align Chinese punctuation with English punctuation in the encoding dictionary of the Chinese BERT-base model and supplement it with extra punctuation frequently used in social media but not previously recorded. As detailed in Table 1, we are concerned with thirteen different types of punctuation, three of which occur in pairs (i.e., parenthesis, book title mark, and double quotation). We categorize all of them into three classes based on their location (i.e., left, right, and middle), which is useful for punctuation-location attention in Section 4.2.1. Afterward, the word sequence can be fully and accurately translated into a token sequence. Using these tokens, the word- and sentence-level representation are more relevant, facilitating the construction of attention mechanisms that assist the classifier in capturing both intra-class similarity and inter-class differences.

4.1.2. Word and sentence representation

Following processing through the stack of 12 transformer layers, the Chinese BERT-base model generates n 768-dimensional vectors to represent the sequence (as depicted in the representation sequence shown in Fig. 2). Thanks to the self-attention architecture, the embedding of 'CLS' aggregates information from every word, which can be utilized for concise representation of the whole sentence. For most downstream tasks, especially classification, the representation of 'CLS' is adequately informative. It has been demonstrated as an effective feature input for various NLP tasks and greatly simplifies the computational complexity of the classifier (Ren et al., 2021; Zhang et al., 2023; de Andrade et al., 2023). We keep this commonly used representation as comparative experiments in Section 5.2.2; nonetheless, this compression loses the original position information of words. PLASA uses the whole representation sequence without fine-tuning for following attention and sentiment polarity classification. Utilizing either average or maximum pooling operation across the whole representation sequence is another efficient compression approach (Reimers and Gurevych, 2019). Due to the incapability of machine-learning classifiers to directly process 3D inputs, pooling becomes essential before converting the 3D inputs into 2D through vector concatenation.

4.2. Collaborative attention module

In recent years, the attention mechanism has emerged as a prominent research topic in both the Computer Vision (CV) and NLP domains, playing a vital role in a wide range of models (Li et al., 2022). Attention can be interpreted as weights, and the mechanism represents the relationships among these weights. It aims to strengthen important features and suppress unimportant ones in both images and language. Transformers have become the mainstream framework for deep learning, thanks to the self-attention architecture that enables parallel computation of internal relationships within input data. Self-attention is a critical application of attention mechanisms. Universal attention mechanisms extend their scope beyond word sequences themselves, as they can utilize additional supporting information (Wang et al., 2023; Liu et al., 2023; Xiao et al., 2022). Considering that individuals often skim through vast volumes of content on social media with only a glance, our minds must rapidly react and capture the essential keywords in comments. To emphasize more meaningful features in comment representation, we introduce the collaborative attention module, which

[4] https://github.com/fxsjy/jieba.

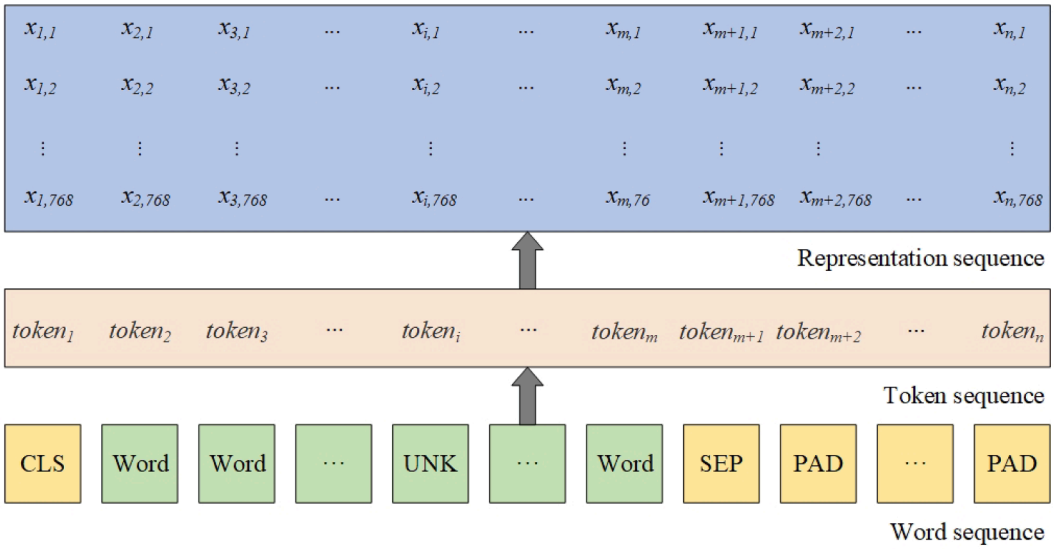

Fig. 2. The representation process of the Chinese BERT-base model.

Table 1
The punctuation of model concerns. R refers to the punctuation located on the sentence' right side; L refers to the punctuation located on the sentence' left side; C refers to the punctuation located on the sentence's middle.

No	Name	Translation	Symbolic	Location	Inclusion
1	逗号	Comma	,	R	√
2	句号	Period	。	R	√
3	问号	Question mark	?	R	√
4	感叹号	Exclamation mark	!	R	√
5	冒号	Colon	:	L	√
6	分号	Semicolon	;	R	√
7	顿号	Enumerated comma	、	M	√
8	波浪号	Tilde	~	R	√
09[a]	左括号	Left parenthesis	(	L	√
09[b]	右括号	Right parenthesis	)	R	√
10[a]	左书名号	Left book title mark	《	L	√
10[b]	右书名号	Right book title mark	》	R	√
11	省略号	Ellipsis	…	R	×
12	破折号	Dash	—	L	×
13_a	左双引号	Left double quotation	“	L	×
13^b	右双引号	Right double quotation	”	R	×

combines punctuation-location and lexicon-query attentions (as described in Fig. 3). These novel attention mechanisms help our model know 'where' and 'what' words should be given more attention (weights), respectively.

4.2.1. Punctuation-location attention

The BERT model has employed position embeddings as a supplement for ensuring the temporal ordering of the text (Devlin et al., 2019). Building on this inspiration, we present relative position embeddings respect to punctuation, which leads to the generation of punctuation-location attention. In the previous Section 4.1.1, we define the scope of punctuation that we are focusing on and categorize their location attributes (i.e., left, middle, and right). Based on human perception, left-location punctuation (e.g., ":") highlights content on its right, right-location punctuation (e.g., " ! " and " ? ") emphasizes content on its left, and middle-location punctuation (e.g. "、") draws attention to content on both sides. Building upon this concept, words positioned closer to punctuation naturally receive more attention from readers.

For the first time, we introduce a hierarchical window $[n, \cdots, 2, 1]$ as an encoding scheme to mark the words adjacent to the right of the left-location punctuation, n is the dimension of this window. Similarly, we operate another hierarchical window $[1, 2, \cdots, n]$ as an encoding scheme to mark the words adjacent to the left of the right-location punctuation. Moreover, we simultaneously adopt two different hierarchical windows $[n, \cdots, 2, 1]$ and $[1, 2, \cdots, n]$ as encoding schemes to mark words on both sides of the middle-location punctuation. When a word simultaneously receives attention from different punctuation, we retain the maximum value of marks. Additionally, if a word doesn't fall within the window of any punctuation, it is assigned an encoding value of 0. After that, these relative position embeddings respect to punctuation are generated ($[l_1, l_2, \cdots, l_i, \cdots, l_n], l_i \in \{0, 1, \cdots, n\}$). Acknowledging that the change in attention might not adhere to a linear trend, it's important to perform a nonlinear variation. Since sine curves have been demonstrated to be suitable for position embedding (Wang et al., 2020), we opt for the $sin\pi/(2n)x$ function to reflect these nonlinear relationships. Here, n corresponds to the dimension of the hierarchical windows, ensuring that the mapping's output remains within the range of [0, 1]. The ultimate punctuation-location attention is given by the softmax activation, as formulated by Eq. (1).

$$attention\_p = softmax\{sin\left(\frac{\pi}{2n}[l_1, l_2, \cdots, l_i, \cdots, l_n]\right)\} * n \tag{1}$$

4.2.2. Lexicon-query attention

Leveraging empirical knowledge from lexicons is another potential approach to enhance the representation of important features. In recent years, with the prevalence of pre-trained models, the lexicon-based approaches have gradually faded out of the research. However, the pre-trained models and lexicon-based approaches are not mutually exclusive, and the potential for a combination of the two lies in the attention mechanism. LIWC is a commonly used lexicon-based tool for text feature analysis. Among the features it extracts, the affective processes, including positive-emotion, negative-emotion, anxiety, anger, and sadness, are essential elements. Therefore, we select the segments of LIWC's lexicon related to affective processes and negations as the foundation for our query lexicon and incorporate a specialized lexicon known as the "Ontology of Emotional Vocabulary".[5] Hence, we can leverage lexicon-query attention to adjust the relationship between sentiment-related keywords and background words in a sentence. Specifically, we query the lexicons for every word in the word sequence and return results of either 0 or 1. Here, 1 signifies that the queried word exists in lexicons, while 0 signifies its absence.

[5] https://ir.dlut.edu.cn/info/1013/1142.htm.

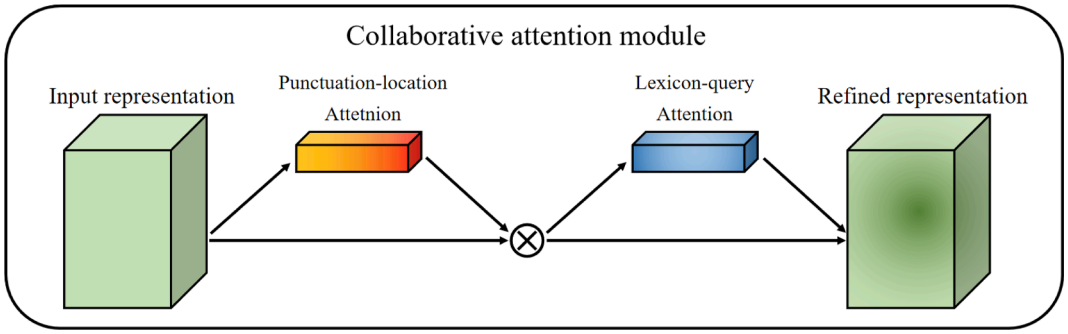

Fig. 3. The overview of the collaborative attention module. The module is divided into two sub-modules, and the representation vector of each comment is sequentially multiplied by two attention mechanisms.

It's worth noting that the target words of our query are derived from traditional Chinese word segmentation, and each character within a word is assigned the same value to map BERT's dimension $[e_1, e, \cdots, e_i, \cdots, e_n]$, $e_i \in \{0, 1\}$. Moreover, this one-hot encoding requires a coefficient θ to amplify the intensity. Finally, the assignment of weights is normalized using the softmax function, and the lexicon-query attention can be calculated by Eq. (2).

$$attention\_p = softmax(\theta^*[e_1, e, \cdots, e_i, \cdots, e_n])^*n \tag{2}$$

Collaborative attention is obtained by multiplying the punctuation-location attention and the lexicon-query attention. The resulting refined representation is referred to as Eq. (3).

$$refined\_r = representation^*attention\_p^*attention\_l \tag{3}$$

4.3. Classification module

The refined representation of each comment is then input into the classification module for sentiment polarity classification. Our research involves a diverse selection of classifiers widely utilized in both industry and academia. These classifiers can be categorized into two groups: machine-learning approaches, including RF, SVM, and XGBoost; and deep-learning approaches, which consist of RNN, and BiLSTM.

5. Datasets and experiment

5.1. Datasets

5.1.1. Data acquisition

Currently, sentiment analysis research in social media is predominantly focused on Twitter, leading to the publicly available datasets being quite uniform (Jurek et al., 2015). China's social media platforms have a plethora of active users whose potential value has not been thoroughly researched. However, there is a lack of labeled datasets suitable for conducting short text sentiment analysis in Chinese social media, especially barrages. In light of this, we utilize three corpora sourced from China's three most popular social media platforms - Sina Weibo, Bilibili, and TikTok - to curate publicly available datasets. In terms of content formats and user demographics, each platform possesses a distinct focus and thrives within its niche domain as follows.

Sina Weibo was launched as early as 2009 and has similarities to Twitter (Zhao et al., 2016). It is primarily used to share instant information about various topics (e.g., entertainment, culture, and education) through images and text. It's worth noting that its users tend to have high levels of cultural and digital literacy (Yin, 2020; Bai and Yu, 2016). Comments on Sina Weibo represent users' self-expression and often contain abundant sentiment information (Zhang and Pentina, 2012; Ling et al., 2020; Wei et al., 2022). As a first corpus, we take the publicly available Weibo_Senti_100k dataset, which contains more than one hundred thousand labeled comments and has been used extensively in previous sentiment analysis (Mu et al., 2021; Zhang et al., 2022). Compared to ultra-short barrages, Weibo comments are relatively longer.

Bilibili is often referred to as the "YouTube of China" and stands out for its unique features, such as ad-free environment and strong connection to anime culture (Zhang and Cassany, 2020). Most importantly, it is seen as a knowledge-sharing platform where users can easily upload their self-produced instructional videos and have unrestricted access to learning content created by others (Li & Zhao, 2023). As a result, Bilibili has fostered a vibrant student community. Despite the modest size and purchasing power of this demographic, they demonstrate a deep reliance on this platform (Li et al., 2023). And, Bilibili is where the barrage comments first caught fire. To focus on sentiment analysis of barrages, given an absence of publicly available datasets fully suited to this task, we used the official open API[6] of Bilibili to collect barrages from the top 100 most popular videos on the platform. There are no specific restrictions on the topics of these videos, which include education, sports, music, games, etc. The data collection was carried out on June 12, 2023.

TikTok has achieved global success owing to its seamless process for creating, sharing, watching, and commenting on short videos, as well as its exceptional recommendation system (Zhang and Liu, 2021). By leveraging data from users' past online behavior, the recommendation system learns their personalized preferences and ensures the delivery of compelling content. Compared to Sina Weibo and Bilibili, TikTok has a broader user base, consisting mainly of more mature individuals with higher levels of consumption. Based on this, TikTok's live-streaming e-commerce marketing has been very successful in China (Barta et al., 2023; Li et al., 2021b). These completed transactions are frequently driven by positive sentiment, and the barrages generated during live-streaming effectively mirror the perceptions and cognitions of the audience. We carefully select two live streams, "East Buy" (Chinese name: 东方甄选) and "Make a Friend" (Chinese name: 交个朋友), whose fans have relatively high levels of cultural and digital literacy, to collect real-time barrages and ensure that the analysis targets are meaningful. From October 22 to November 1, 2022, we used the official open API[7] of TikTok for persistent data gathering. Coinciding with the COVID-19 pandemic, people's sentiments fluctuated.

No personal information is identified or stored during the crawling process, and the collected data is solely intended for academic research purposes.

5.1.2. Data preprocessing

Upon acquiring our desired corpora, we performed the following data preprocessing steps to form the final datasets for model training.

[6] https://openhome.bilibili.com.
[7] https://developer.open-douyin.com.

- Empty and Duplicated Items Deletion: When capturing corpora directly through scripts, it is common to encounter empty and duplicate items. Especially in the context of videos and live-streaming, preceding barrages can strongly impact other viewers, and the existence of a one-click copy and publish feature offered by platforms contributes to the spread of repetitive comments. Therefore, deleting empty and duplicated items in the initial step can effectively reduce the workload of subsequent steps.
- Noisy Text Removal: In this step, the noisy patterns of comments or barrages are removed from the text, including emojis (e.g., 😭[哭], 👍[赞]), hashtags (e.g., #keyword#), user names (e.g., @王**), and URLs. Emojis express explicit emotional tendencies, allowing us to sense the sentiment of a comment without having to read the words. However, there is a possibility of conflict between the use of emojis and the semantics of words. Also, hashtags, user names, URLs, and other noisy patterns that can cause misunderstandings should be removed.
- System-generated Content Filtering: The purpose of this step, designed for the barrages, is to filter out any system-generated content. It's worth noting that video producers, live streaming hosts, or platforms can insert pre-set content into users' barrage sequences at any point to guide user behavior. The language used in such system-generated content is similar, featuring terms like "福袋" (translation: lucky bag) and "一键三连" (translation: like and subscribe). We thoroughly retrieved each barrage to ensure that all items were the result of human typing.
- Information-poor Comments Elimination: This step aims to guarantee that all comments contain essential linguistic components. The principle is that each comment must consist of at least two words. Punctuation and numbers are considered as words, but they cannot be the only element of a comment. To achieve this, we employ the jieba library for word segmentation and part-of-speech (POS) tagging before counting.

As our primary research objective is to explore the role of punctuation and lexicon on user expressions, we focus exclusively on texts with punctuation. It is a common practice to exclude sentiment-absent samples before data labeling, which can significantly reduce manual effort. However, when applied to social media short text, this process is challenging due to the presence of incomplete emotion words. In particular, traditional lexicons often overlook abbreviations and acronyms. Additionally, contemporary and platform-specific words may also not be included in lexicons (Alshaabi et al., 2022), such as "火钳刘明" (meaning: outstanding video) and "静默" (meaning: isolation at home). Punctuation can also influence the sentiment of words. For instance, it can give neutral words a sentiment tendency and reverse the sentiment polarity of certain words. These details should be professionally evaluated. Several studies have employed rule-based or machine-learning methods to generate labeled datasets based on features such as emoticons, emojis, hashtags, emotional words, negations, and punctuation (Upadhyaya et al., 2023; Ptaszynski et al., 2014). However, the labeling results can be biased due to the potential conflict between these features and the semantics of the text. We observe this conflict phenomenon in the Weibo_Senti_100k dataset, which requires manual re-labeling. Since the Weibo_Senti_100k has more than one hundred thousand comments, we use random down-sampling to balance its quantitative relationship with the other two datasets.

- Data labeling: We engaged three annotators to label data, each with at least two years of experience in social media-related work or research. Since it was a binary classification task, two rounds of labeling were conducted according to the voting principle. In the first round, three annotators were asked to label the datasets independently, discarding some of the sentiment-absent samples or whose polarity was difficult to determine (number of labels < 2). In the second round, the comments that resulted in inconsistent labels were collected and sent back to the annotators for re-evaluation, taking into account the results of other annotators. After that, the results of Fleiss' Kappa (Kabir et al., 2023) on each of the datasets indicated a high level of inter-rater consistency, with all exceeding 0.8. The final labeling result was determined by the polarity that received more votes (number of votes > 1).

Finally, we obtained three datasets, namely SentiTikTok-2023, SentiBilibili-2023, and SentiWeibo-2023. In Table 2, we present some example datasets. At the same time, we have made our datasets openly available without any reservations in order to ensure the reproducibility of experiments and promote advancement in the field.

5.1.3. Dataset statistics

The statistics of the three datasets are illustrated in Table 3. It can be observed that the length of barrages in TikTok's live streams is shorter compared to the barrages in Bilibili's videos. This difference can be attributed to the timeliness of communication in live streaming on TikTok, while users can organize their language by pausing the videos on Bilibili. On the contrary, Sina Weibo, as a more older social media platform, results in lengthier comments due to content fragmentation and a lack of immediacy. Another important point to note is that, despite the variation in text length across different social media platforms, the proportions of punctuation respect to the texts are similar. This similarity suggests a sharing user behavior in language. Moreover, the consistent percentage (approximately 13 %) indicates that punctuation plays a vital role in expression and has the potential to influence NLP research.

We employ the LIWC program to perform a statistical analysis of LIWC features in three datasets, facilitating a better understanding of the characterization of social media short text. Due to the concise nature of the text and the need to count over 100 categories, the LIWC output is highly sparse. Our objective is to identify significant lexicon features and measure the distinctions between negative and positive samples. To accomplish this, we compute the Shannon entropy (Shannon, 1948) for the positive and negative segments in each dataset using Eq. (4). Based on the results, various critical features are illustrated in Fig. 4.

$$H(X) = -\sum_{i} P(X_i) log_2 P(X_i) \tag{4}$$

Figs. 4(a) and 4(b) present the entropy of ten LIWC features in the three datasets, comparing the positive and negative classes. The entropy of the positive emotion (abbr. PosEmo) and the entropy of the

Table 2
The examples of three datasets, including original texts, translations, and labels.

Dataset	Example	Translation	Label
SentiTikTok-2023	今天这身确实帅！	You look really cool in this outfit today!	1
	离老百姓消费越来越远。	It's becoming increasingly unaffordable for the average person.	0
SentiBilibili-2023	古人的聪明真的是让人惊叹！	The wisdom of the ancients is truly astonishing!	1
	剪辑有点土。。。好尬的音效	The editing is a bit crude… and the sound effects are awkward.	0
SentiWeibo-2023	享受美食就是享受快乐！谢谢喜欢我们家的三样菜哦！记得开心每天啦！	Enjoying good food is enjoying happiness! Thank you for liking our three dishes! Remember to be happy every day!	1
	足足一小时, 报社楼下都没车位, 这可怎么办哦？我Hold不住了！！	For a whole hour, there were no parking spots under the newspaper building. What should I do now? I can't handle it anymore!!	0

Table 3
The statistics of the three datasets. Items refer to the total number of comments; Max. L indicates the maximum text length; Avg. L indicates the average text length; Avg. P is the average number of punctuation; Pct. P is the proportion of punctuation in the text.

Dataset	Items	Max. L	Avg. L	Avg. P	Pct. P
SentiTikTok-2023	4613	28	9.33	1.17	12.52 %
SentiBilibili-2023	7755	94	15.24	1.93	12.68 %
SentiWeibo-2023	5614	159	42.53	5.33	12.53 %

exclamation mark (abbr. Exclam) are relatively higher in the positive class, while the entropy of the negative emotion (abbr. NegEmo) and question mark (abbr. QMark) are more prominent in the negative class. However, it is important to note that the "Exclam" is not solely attributed to the positive class as it serves as an emotional reinforcement. Therefore, its entropy is not low in the negative class. The findings from LIWC provide us with valuable insights that sentiment-related words and punctuation have a significant influence on sentiment analysis.

According to the maximum length of items and BERT's encoding mode, the samples in SentiTikTok-2023 involve 30 tokens, those in SentiBilibili-2023 involve 96 tokens, and those in SentiWeibo-2023 involve 161 tokens (Max. L + 2). While most items in the datasets are shorter than their respective maximum lengths, padding has only a minimal impact on the model training.

5.2. Experiment

5.2.1. Experimental setup

Our experiments are carried out in the Python 3.9 environment with the necessary libraries. The deep learning aspect is implemented using the PyTorch framework and supported by an Nvidia 8G GPU. PLASA uses the Chinese BERT-base model for word- and sentence-level representation without fine-tuning due to incoherence and repetitive contexts, which can lead to underfitting or overfitting. The three datasets were randomized 8:1:1 into train, validate, and test sets. The train and validate sets are used for model training and parameter optimization, whereas the test set is for comparison and ablation experiments.

When the classifier is the RNN or BiLSTM network, we employed the AdamW optimizer (Martyn and Kadziński, 2023) and the cross-entropy loss function (Meng et al., 2023) for model training with a learning rate of 0.0005. And when the classifier is RF, SVM, or XGBoost, we follow several tricks, including Grid search (Chen et al., 2022) and Bayesian optimization (Jain et al., 2023), to efficiently filter suitable hyperparameters for different models and ensure optimal performance. In terms of evaluation metrics, macro-F1, weighted-F1, and micro-F1 are used. It should be noted that accuracy and micro-F1 are equivalent in binary classification tasks (Grandini et al., 2020).

5.2.2. Experimental results

Table 4 shows the overall performance of the commonly used models (BERT_CLS), the baseline models (Baseline), and the hybrid models we proposed (PLASA) on three datasets. The results show that our model PLASA, which integrates the whole BERT representation, the collaborative attention module, and the BiLSTM network, outperforms all other models across datasets and metrics. Overall, the performance of all models decreases as the comment length increases across datasets. It is worth noting that the representation of commonly used models compresses each text of varying lengths into a 768-dimensional vector, which still provides excellent semantic information for the classification task. In addition, within these commonly used models, the machine-learning classifiers also demonstrate their remarkable capabilities, especially when the performance of the FC layer on SentiWeibo-2023 drops sharply, and the machine-learning classifiers remain comparatively stable.

To further investigate the relationships between sequences and take advantage of our proposed punctuation-location and lexicon-query attentions, we introduce the whole representation sequences in both the baseline and PLASA models. In our experiments, we observe that machine-learning classifiers are not adept at handling sequential data. Therefore, for classification models, achieving better performance does not depend solely on having more input features, but rather on how efficiently features are extracted. In the baseline models, the deep-learning classification networks RNN and BiLSTM appear to be better at processing sequences.

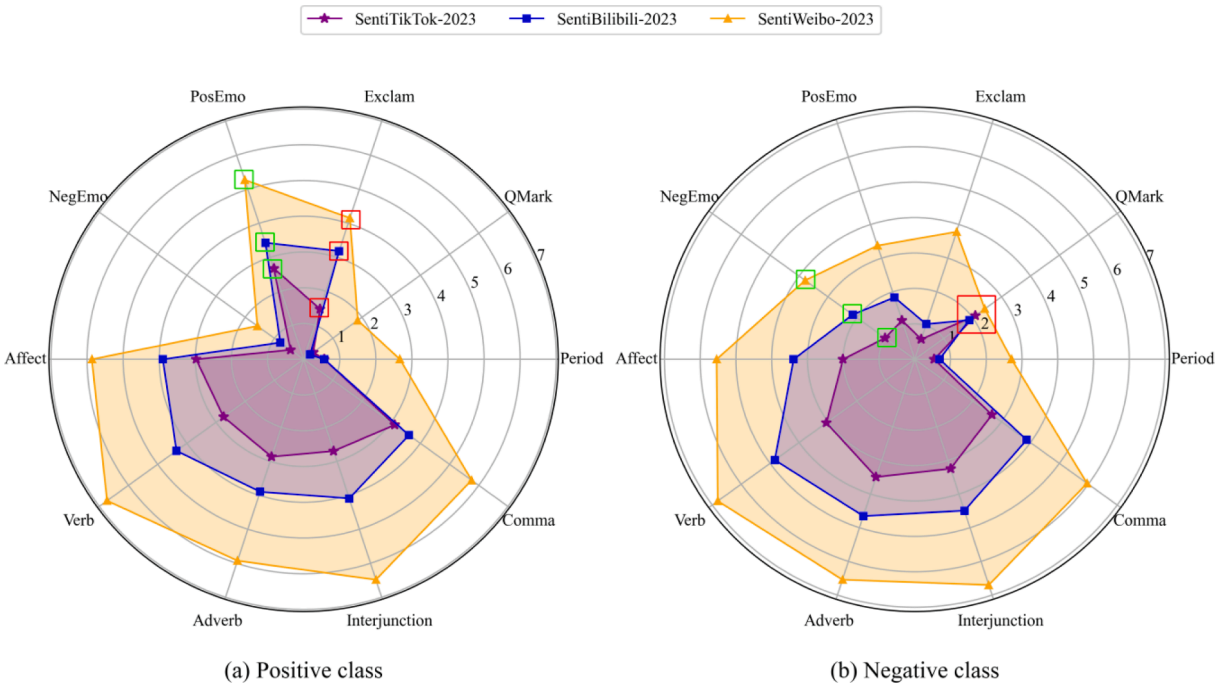

Fig. 4. The LIWC features of datasets. In the radar charts, green boxes indicate the differences in emotional word usage, while red boxes highlight the differences in punctuation usage. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
The comparison performance of different models on datasets. The best results are highlighted in bold.

	Model	SentiTikTok-2023			SentiBilibili-2023			SentiWeibo-2023		
		macro-F1	weighted-F1	micro-F1	macro-F1	weighted-F1	micro-F1	macro-F1	weighted-F1	micro-F1
BERT_CLS	BERT_CLS_RF	0.8639	0.8838	0.8874	0.8689	0.8690	0.8690	0.8452	0.8452	0.8452
	BERT_CLS_SVM	0.8575	0.8736	0.8723	0.8664	0.8664	0.8664	0.8681	0.8683	0.8683
	BERT_CLS_XGBoost	0.8918	0.8734	0.8903	0.8809	0.8809	0.8810	0.8719	0.8719	0.8719
	BERT_CLS_FC	0.8868	0.8853	0.8727	0.8716	0.8717	0.8717	0.8334	0.8340	0.8345
Baseline	BERT_RF	0.8001	0.8313	0.8398	0.7665	0.7663	0.7672	0.7562	0.7563	0.7562
	BERT_SVM	0.8328	0.8531	0.8528	0.7736	0.7737	0.7738	0.7687	0.7687	0.7687
	BERT_XGBoost	0.8280	0.8519	0.8550	0.7920	0.7921	0.7923	0.7705	0.7705	0.7705
	BERT_RNN	0.8402	0.8653	0.8723	0.8470	0.8468	0.8468	0.8301	0.8307	0.8310
	BERT_BiLSTM	0.8849	0.9007	0.9026	0.8836	0.8837	0.8839	0.8350	0.8360	0.8381
PLASA	PLA_RF	0.8023	0.8334	0.8420	0.7653	0.7656	0.7672	0.7615	0.7616	0.7616
	PLA_SVM	0.8323	0.8538	0.8550	0.7591	0.7592	0.7593	0.7292	0.7295	0.7295
	PLA_XGBoost	0.8463	0.8675	0.8701	0.8174	0.8174	0.8175	0.7722	0.7723	0.7722
	PLA_RNN	0.8872	0.8909	0.8896	0.8743	0.8743	0.8743	0.8505	0.8505	0.8505
	PLA_BiLSTM	**0.9392**	**0.9398**	**0.9394**	**0.9032**	**0.9035**	**0.9034**	**0.8879**	**0.8879**	**0.8879**

In order for our PLASA to extract the most relevant and useful features, we incorporate punctuation-location and lexicon-query attentions (PLA) to regulate the representation sequences. PLASA with RNN or BiLSTM shows a significant performance improvement compared to the corresponding baseline model. In addition, the micro-F1 of the best PLASA model (PLA_BiLSTM) is improved by 3.68 %, 1.95 %, and 4.98 % on three datasets, respectively. This demonstrates the sophistication of our hybrid model and the effectiveness of the collaborative attention module. Furthermore, our research narrows the model performance gap between short-text and long-text datasets to some extent by incorporating punctuation information and lexicon knowledge.

To further demonstrate PLASA's superiority, we compare it with the following SOTA models on three datasets.

Topic-level Model (Pathak et al., 2021): The model extracts topics from short social media texts and then incorporates a topic-level attention mechanism with an LSTM for sentiment analysis, utilizing Word2-Vec and Glove embeddings.

SLCABG (Yang et al., 2020): The SLCABG model combines a sentiment lexicon with CNN and BiGRU to enhance the features extracted from BERT embeddings.

RoBERTa (Liu et al., 2019): A refined version of BERT, RoBERTa extends the training process with larger datasets and more robust training techniques. It employs dynamic masking and eliminates the Next Sentence Prediction (NSP) task, significantly enhancing its ability to understand the bidirectional context in text.

XLNet (Yang et al., 2019): Merging autoregressive and autoencoding techniques, XLNet utilizes a permutation-based training approach, enhancing its ability to understand the bidirectional context. It incorporates a two-stream attention mechanism, improving flexibility in text processing.

ERNIE-3.0 (Sun et al., 2021): Capitalizing on knowledge from both structured and unstructured data, ERNIE 3.0 substantially improves its language understanding and generation capabilities across a wide range of NLP applications.

The observations from Table 5 indicate that PLASA surpasses existing SOTA models. Owing to the Topic-level Model's reliance on static representations, it exhibits the lowest accuracy among these models. Furthermore, the enhancement effect of the topic-level attention mechanism is suboptimal for sentiment analysis on short texts, specifically barrages. Although the SLCABG model also utilizes a lexicon to enhance BERT representations, it does not take into account the role of punctuation. The accuracy of recent transformer-based sentiment analysis approaches, though innovative, does not surpass PLASA, and these approaches entail additional computational overhead. However, ERNIE-3.0 demonstrates enhanced efficacy in Chinese sentiment analysis, positioning it as a viable alternative for the representation module in our model for specific tasks.

5.2.3. Ablation study

To ascertain the significance of each unit within the attention module, we conduct an ablation study by splitting the punctuation-location attention and the lexicon-query attention. This allowed us to clarify the individual contributions of each part in the best PLASA model. Two independent experiments were performed: one with the exclusion of punctuation-location attention (PLASA-PA) and the other with the removal of lexicon-query attention (PLASA-LA) in the best model. Table 6 describes the performance of two ablation models alongside PLASA and baseline. As expected by the experimental design, the performance of the two variant models lies between PLASA and the baseline model on all datasets.

The results of the ablation study show the rationality and indispensability of each component. We employ two distinct attention mechanisms, both of which lead to performance improvements in the model. The effects of both attention mechanisms are depicted in Fig. 5. It is evident that the effect of punctuation-location attention is most pronounced on the SentiTikTok-2023 dataset, but this effect diminishes as the text length increases. In longer texts on SentiWeibo-2023, where information is enriched and there is more noise, the role of lexicon-query attention becomes critical. It strengthens the important representations for sentiment classification, helping to improve the performance of the

Table 5
Comparison with existing SOTA models for short text sentiment analysis on social media.

Model	SentiTikTok-2023			SentiBilibili-2023			SentiWeibo-2023		
	macro-F1	weighted-F1	micro-F1	macro-F1	weighted-F1	micro-F1	macro-F1	weighted-F1	micro-F1
Topic-level Model	0.7814	0.8014	0.8049	0.6762	0.6782	0.6949	0.6298	0.6297	0.6351
SLCABG	0.8923	0.8930	0.8931	0.8728	0.8754	0.8742	0.8478	0.8476	0.8848
RoBERTa	0.8033	0.8339	0.8420	0.7111	0.7109	0.7116	0.6799	0.6799	0.6799
XLNet	0.9000	0.9118	0.9113	0.8611	0.8611	0.8611	0.8608	0.8610	0.8612
ERNIE-3.0	0.9142	0.9256	0.9264	0.8808	0.8810	0.8823	0.8559	0.8559	0.8559
PLASA	**0.9392**	**0.9398**	**0.9394**	**0.9032**	**0.9035**	**0.9034**	**0.8879**	**0.8879**	**0.8879**

Table 6
The impact of different attentions on datasets.

Model	SentiTikTok-2023			SentiBilibili-2023			SentiWeibo-2023		
	macro-F1	weighted-F1	micro-F1	macro-F1	weighted-F1	micro-F1	macro-F1	weighted-F1	micro-F1
PLASA	0.9392	0.9398	0.9394	0.9032	0.9035	0.9034	0.8879	0.8879	0.8879
PLASA-LA	0.9103	0.9206	0.9199	0.8981	0.8981	0.8981	0.8447	0.8451	0.8452
PLASA-PA	0.9085	0.9198	0.9199	0.9008	0.9008	0.9008	0.8754	0.8755	0.8754
Baseline	0.8849	0.9007	0.9026	0.8836	0.8837	0.8839	0.8350	0.8360	0.8381

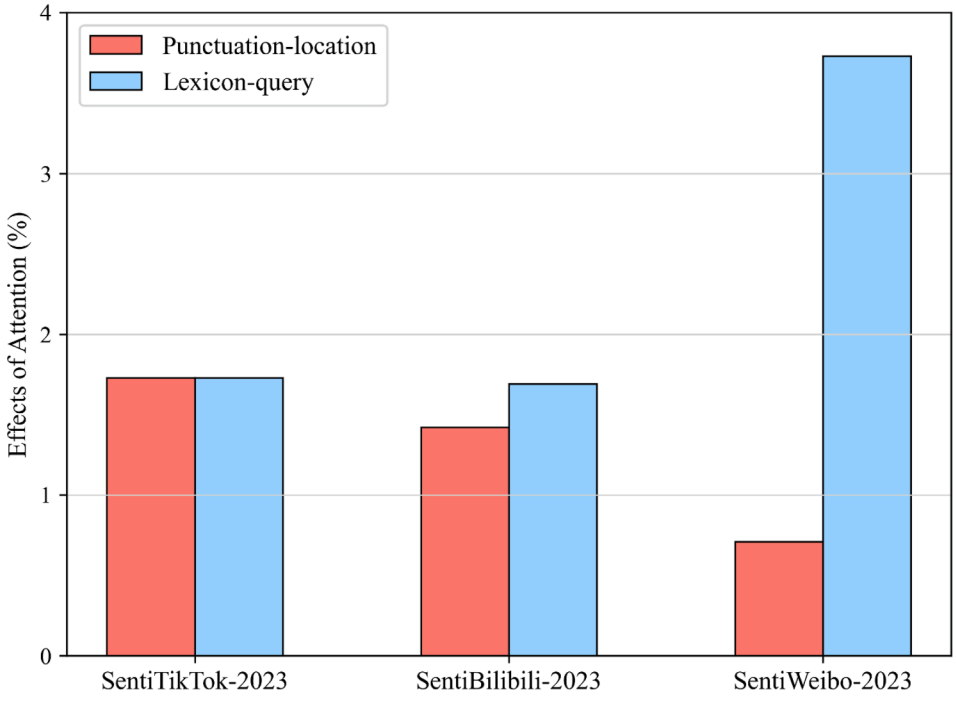

Fig. 5. The effects of punctuation-location and lexicon-query attentions.

model while suppressing the influence of noise. Importantly, the effects of these two components are neither repetitive nor additive, and each part plays a distinct and critical role, collaboratively contributing to the outstanding performance of the PLASA. Finally, in a small subset of the SentiWeibo-2023 test sets, we visualize the collaborative attention in Fig. 6.

5.2.4. *Parameter sensitivity*

The proposed punctuation-location and lexicon-query attentions require setting two hyper-parameters before model training, the dimension of window (n) and the coefficient of intensity (θ). On the one hand, in punctuation-location attention (refer to Section 4.2.1), we search for appropriate dimension n within set $\{3, 4, 5, 6, 7\}$, and simultaneously vary the value of n in $sin\pi/(2n)x$. On the other hand, in lexicon-query attention (refer to Section 4.2.2), we should control the coefficient θ in $\{0.5, 0.75, 1, 1.25\}$ so that background words do not decay to insignificance when sentiment-related words gain more weight. Hence, to achieve an optimal trade-off between these two components, we conduct parameter sensitivity experiments on three datasets. We evaluate the results using the metric of micro-F1, and the results are shown in Fig. 7. We first set θ to 1 to compare the different values of n, and then we set n to 5 to explore various values of θ. The effect of different parameter combinations slightly fluctuates across datasets, but they already outperform the baseline model. In summary, the combination of parameters $n = 5$ and $\theta = 1$ exhibits stability and relatively great performance, serving as an initial value for the collaborative attention module.

6. Discussion and implications

Sentiment analysis needs to keep up with the diversity of language forms. Previous studies have not thoroughly investigated Chinese social media platforms, especially when it comes to sentiment analysis of barrages. The primary challenge is that ultra-short barrages might lack the necessary semantic information for effective sentiment analysis, whereas longer comments could introduce more noise complicating the determination of sentiment polarity. Recent research for social media sentiment analysis heavily relies on pre-trained models, ignoring the empirical knowledge of lexicon-based approaches. Although punctuation is an important component of language expression and its influence on the sentiment of comments is acknowledged, there is a lack of a combination between punctuation usage and word representation. In this background, we propose a hybrid model called PLASA integrating punctuation information and lexicon knowledge. The noteworthy results have implications for both theoretical and practical aspects.

In terms of the theoretical aspect, we first observe that punctuation consistently makes up approximately 13 % of comments with varying lengths, indicating the vital role of punctuation in user language expression. And the lexicon features between the positive and negative segments in the datasets differ significantly. These findings signify the importance of exploring the potential of punctuation and lexicon usage to enhance sentiment analysis performance. In PLASA, we introduce punctuation-location and lexicon-query attentions to improve the performance of sentiment analysis, taking into consideration human perception and cognition. Theoretically, through comparison and ablation experiments, we have demonstrated the sophistication of PLASA

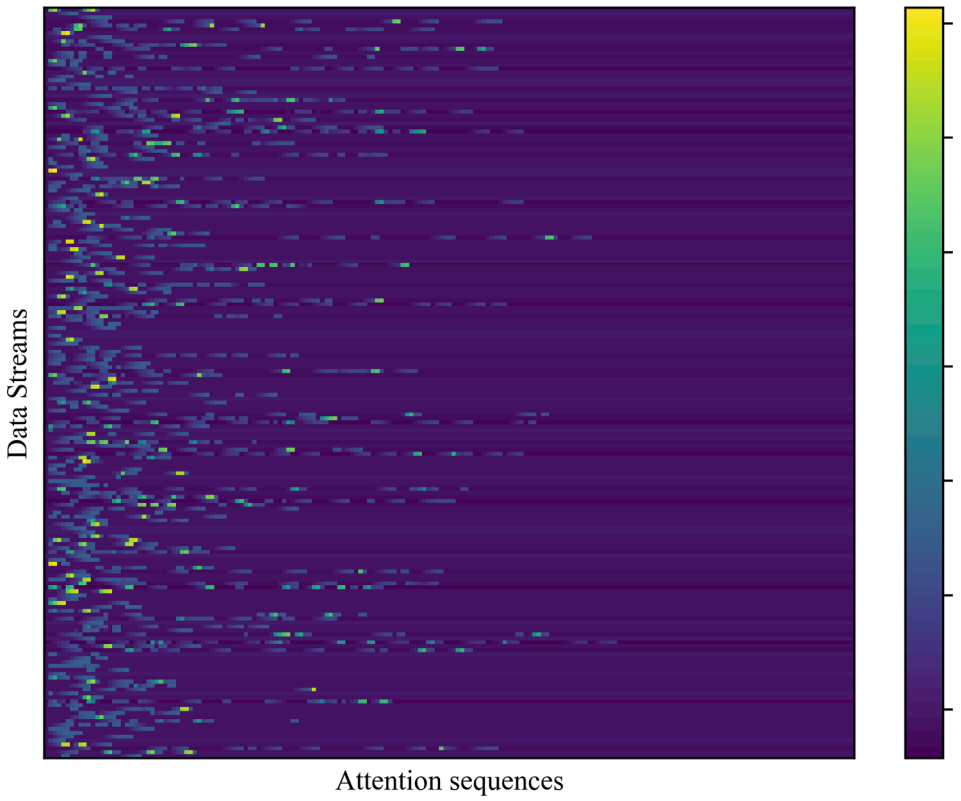

Fig. 6. The visualization of collaborative attentions. The horizontal axis represents the sequence of comment data streams, while the vertical axis represents the dimension of the embedding weights. Brighter areas signify greater weights than darker areas.

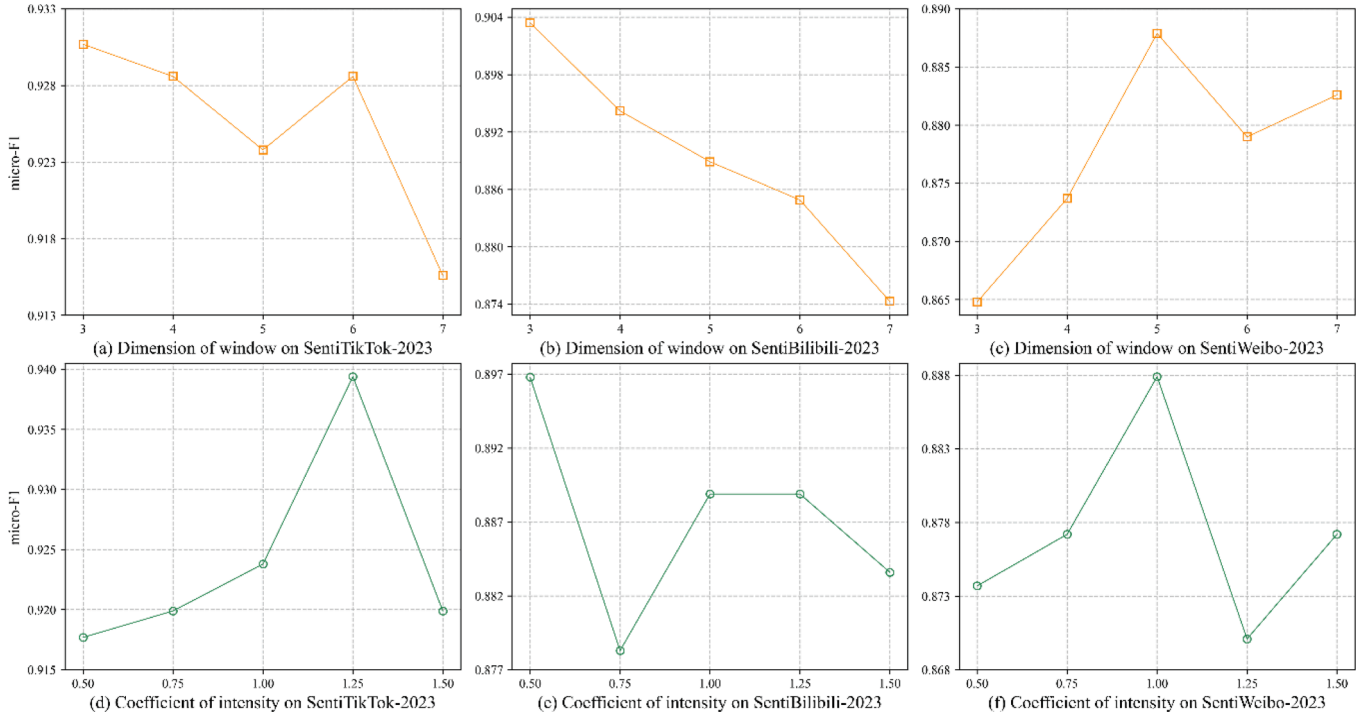

Fig. 7. The sensitivity analysis of the dimension of window and coefficient of intensity on datasets.

and the effectiveness of the collaborative attention module.

In terms of practical aspects, we have curated three distinct datasets of Chinese social media comments that have received little attention in previous research. The investigation of user groups with high levels of cultural and digital literacy is meaningful due to their intricate language expression. Our publicly available datasets may promote advancement in the field. Additionally, the success of PLASA underscores the value of empirical knowledge derived from traditional methods, enhancing the

interpretability of deep learning models. In real-world applications, the punctuation-location and lexicon-query attentions can be easily inserted into various NLP tasks either collaboratively or independently, as required. Given that sentiment analysis is an essential means of measuring user experience, our methodology can benefit content creators, platforms, technology service companies, and government agencies in management practice. Lastly, it provides a solid foundation for future social media research on live streaming and short videos.

7. Conclusion and future work

This paper proposes PLASA, a hybrid model designed for classifying social media comments into two polarized categories: positive and negative. PLASA introduces a collaborative attention module, incorporating position information respect to punctuation and empirical knowledge from lexicons to determine 'where' and 'what' words to be emphasized in representation. We have curated two novel barrage datasets in full flow, which involve data acquisition, preprocessing, and sentiment polarity labeling. Additionally, we have rectified a benchmark dataset that originally contained automatically generated labels. The comparison and ablation experiments demonstrate the superior performance of PLASA, while the collaborative attention module showcases great promise. Our study represents a significant step forward to ultra-short texts (barrages) while simultaneously addressing the problem of high noise informal language for social media sentiment analysis. The datasets and model we present are valuable for text classification tasks and subsequent research works on live streams and short videos. Furthermore, our research can offer benefits to content creators, platforms, technology service companies, and government agencies in practice.

Despite PLASA performing well on datasets from three different social media platforms, the scale of these datasets is relatively small. Future research should incorporate both larger datasets and more extended texts, further validating PLASA's robustness and generalizability across different contexts. Our research exemplifies a successful integration of state-of-the-art technology and traditional methods in the field of NLP. However, numerous potentials warrant further investigation, such as applying our attention mechanism with aspect-level sentiment analysis (Ahmed et al., 2023), exploring additional management theory as support, and utilizing advanced AI methods to optimize the model. Since sentiment analysis plays a crucial role in measuring user experience at both the individual and group levels, real-time sentiment analysis can elucidate a causal relationship with social media content (e.g., visual and auditory information). In the future, our research efforts will be dedicated to multimodal sentiment analysis (Gandhi et al., 2023) from diverse sources.

CRediT authorship contribution statement

Zhenyu Li: Conceptualization, Methodology, Investigation, Data curation, Software, Visualization, Writing – original draft, Writing – review & editing. **Zongfeng Zou:** Conceptualization, Supervision, Project administration, Funding acquisition, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets and model implementation codes are available over this repository: https://github.com/Sometimesrains/Barrage.

Acknowledgements

This study was supported by the National Social Science Fund of China [21ZDA105] and the Shanghai Municipal Science and Technology Major Project [22692117400].

References

Ahmed, T., Ivan, S., Kabir, M., Mahmud, H., Hasan, K., 2022. Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. Soc. Netw. Anal. Min. 12 (1), 99.
Ahmed, K., Nadeem, M.I., Zheng, Z., Li, D., Ullah, I., Assam, M., Mohamed, H.G., 2023. Breaking down linguistic complexities: a structured approach to aspect-based sentiment analysis. J. King Saud Univ.-Comput. Inform. Sci. 35 (8), 101651.
Al Amrani, Y., Lazaar, M., El Kadiri, K.E., 2018. Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Comput. Sci. 127, 511–520.
Alessia, D., Ferri, F., Grifoni, P., Guzzo, T., 2015. Approaches, tools and applications for sentiment analysis implementation. Int. J. Comput. Appl. 125 (3).
Alshaabi, T., Van Oort, C.M., Fudolig, M.I., Arnold, M.V., Danforth, C.M., Dodds, P.S., 2022. Augmenting semantic lexicons using word embeddings and transfer learning. Front. Artificial Intell. 4, 783778.
Alwehaibi, A., Bikdash, M., Albogmi, M., Roy, K., 2022. A study of the performance of embedding methods for Arabic short-text sentiment analysis using deep learning approaches. J. King Saud Univ.-Comput. Inform. Sci. 34 (8), 6140–6149.
Bai, C., Duan, Y., Liu, C., Qiu, L., 2022. International taxation sentiment and COVID-19 crisis. Res. Int. Bus. Financ. 63, 101783.
Bai, H., Yu, G., 2016. A Weibo-based approach to disaster informatics: incidents monitor in post-disaster situation via Weibo text negative sentiment analysis. Nat. Hazards 83 (2), 1177–1196.
Balahur, A., Perea-Ortega, J.M., 2015. Sentiment analysis system adaptation for multilingual processing: the case of tweets. Inf. Process. Manag. 51 (4), 547–556.
Barta, S., Belanche, D., Fernández, A., Flavián, M., 2023. Influencer marketing on TikTok: the effectiveness of humor and followers' hedonic experience. J. Retail. Consum. Serv. 70, 103149.
Bonta, V., Kumaresh, N., Janardhan, N., 2019. A comprehensive study on lexicon based approaches for sentiment analysis. Asian J. Comput. Sci. Technol. 8 (S2), 1–6.
Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., Hassanien, A.E., 2020. Sentiment analysis of COVID-19 tweets by deep learning classifiers—A study to show how popularity is affecting accuracy in social media. Appl. Soft Comput. 97, 106754.
Chen, Z.T., 2020. Slice of life in a live and wired masquerade: playful prosumption as identity work and performance in an identity college Bilibili. Global Media China 5 (3), 319–337.
Chen, H., Zhang, Z., Yin, W., Zhao, C., Wang, F., Li, Y., 2022. A study on depth classification of defects by machine learning based on hyper-parameter search. Measurement 189, 110660.
Chen, G., Zhou, S., Zhi, T., 2019. Viewing mechanism of lonely audience: evidence from an eye movement experiment on barrage video. Comput. Human Behav. 101, 327–333.
Chiarello, F., Bonaccorsi, A., Fantoni, G., 2020. Technical sentiment analysis. measuring advantages and drawbacks of new products using social media. Comput. Ind. 123, 103299.
Choi, D., Hwang, M., Kim, J., Ko, B., Kim, P., 2014. Tracing trending topics by analyzing the sentiment status of tweets. Comput. Sci. Inf. Syst. 11 (1), 157–169.
Crossley, S.A., Kyle, K., McNamara, D.S., 2017. Sentiment analysis and social cognition engine (SEANCE): an automatic tool for sentiment, social cognition, and social-order analysis. Behav. Res. Methods 49, 803–821.
Cureg, M. Q., De La Cruz, J. A. D., Solomon, J. C. A., Saharkhiz, A. T., Balan, A. K. D., & Samonte, M. J. C. (2019, March). Sentiment analysis on tweets with punctuations, emoticons, and negations. In: Proceedings of the 2nd International Conference on Information Science and Systems, pp. 266-270.
Dangi, D., Dixit, D.K., Bhagat, A., 2022. Sentiment analysis of COVID-19 social media data through machine learning. Multimed. Tools Appl. 81 (29), 42261–42283.
de Andrade, C.M., Belém, F.M., Cunha, W., França, C., Viegas, F., Rocha, L., Gonçalves, M.A., 2023. On the class separability of contextual embeddings representations–or "the classifier does not matter when the (text) representation is so good!". Inf. Process. Manag. 60 (4), 103336.
del Pilar Salas-Zárate, M., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., Alor-Hernández, G., 2014. A study on LIWC categories for opinion mining in Spanish reviews. J. Inf. Sci. 40 (6), 749–760.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, Vol. 1, p. 2. https://doi.org/10.18653/v1/N19-1423.
Dhaoui, C., Webster, C.M., Tan, L.P., 2017. Social media sentiment analysis: lexicon versus machine learning. J. Consum. Mark. 34 (6), 480–488.
Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A., 2023. Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion 91, 424–444.
Gaston, J., Narayanan, M., Dozier, G., Cothran, D. L., Arms-Chavez, C., Rossi, M., ... & Xu, J. (2018, November). Authorship attribution via evolutionary hybridization of sentiment analysis, LIWC, and topic modeling features. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 933-940). IEEE.

Geng, S., Niu, B., Feng, Y., Huang, M., 2020. Understanding the focal points and sentiment of learners in MOOC reviews: a machine learning and SC-LIWC-based approach. Br. J. Educ. Technol. 51 (5), 1785–1803.
Govindan, V., Balakrishnan, V., 2022. A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection. J. King Saud Univ.-Comput. Inform. Sci. 34 (8), 5110–5120.
Grandini, M., Bagli, E., Visani, G., 2020. Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.
Hama Aziz, R.H., Dimililer, N., 2021. SentiXGboost: enhanced sentiment analysis in social media posts with ensemble XGBoost classifier. J. Chin. Inst. Eng. 44 (6), 562–572.
Hao, X., Xu, S., Zhang, X., 2019. Barrage participation and feedback in travel reality shows: the effects of media on destination image among generation Y. J. Destin. Mark. Manag. 12, 27–36.
He, L., Yin, T., Zheng, K., 2022. They May Not Work! An evaluation of eleven sentiment analysis tools on seven social media datasets. J. Biomed. Inform. 132, 104142.
Jain, D., Borah, M.D., Biswas, A., 2023. Bayesian optimization based score fusion of linguistic approaches for improving legal document summarization. Knowl.-Based Syst. 264, 110336.
Jurek, A., Mulvenna, M.D., Bi, Y., 2015. Improved lexicon-based sentiment analysis for social media analytics. Security Informatics 4 (1), 1–13.
Kabir, M., Ahmed, T., Hasan, M.B., Laskar, M.T.R., Joarder, T.K., Mahmud, H., Hasan, K., 2023. DEPTWEET: a typology for social media texts to detect depression severities. Comput. Human Behav. 139, 107503.
Koukaras, P., Tjortjis, C., Rousidis, D., 2020. Social media types: introducing a data driven taxonomy. Computing 102 (1), 295–340.
Li, M., Chen, L., Zhao, J., Li, Q., 2021a. Sentiment analysis of Chinese stock reviews based on BERT model. Appl. Intell. 51, 5016–5024.
Li, Y., Li, X., Cai, J., 2021b. How attachment affects user stickiness on live streaming platforms: a socio-technical approach perspective. J. Retail. Consum. Serv. 60, 102478.
Li, Z., Song, J., Qiao, K., Li, C., Zhang, Y., Li, Z., 2022. Research on efficient feature extraction: improving YOLOv5 backbone for facial expression detection in live streaming scenes. Front. Comput. Neurosci. 16, 980063.
Li, L., Zhang, J., An, X., 2023. Using social media for efficient brand marketing: an evaluation of Chinese universities using Bilibili. Socioecon. Plann. Sci., 101645
Li, B., Zhao, J., 2023. Visual-audio correspondence and its effect on video tipping: evidence from Bilibili vlogs. Inf. Process. Manag. 60 (3), 103347.
Ling, M., Chen, Q., Sun, Q., Jia, Y., 2020. Hybrid neural network for Sina Weibo sentiment analysis. IEEE Trans. Comput. Social Syst. 7 (4), 983–990.
Liu, Y., Bi, J.W., Fan, Z.P., 2017. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. Inf. Sci. 394, 38–52.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
Liu, X., Tang, H., Zhao, J., Dou, Q., Lu, M., 2023. TCAMixer: a lightweight mixer based on a novel triple concepts attention mechanism for NLP. Eng. Appl. Artif. Intel. 123, 106471.
Martineau, J., & Finin, T., 2009, March. Delta tfidf: An improved feature space for sentiment analysis. In: Proceedings of the International AAAI Conference on Web and Social Media (Vol. 3, No. 1, pp. 258-261).
Martyn, K., Kadziński, M., 2023. Deep preference learning for multiple criteria decision analysis. Eur. J. Oper. Res. 305 (2), 781–805.
Melton, C.A., White, B.M., Davis, R.L., Bednarczyk, R.A., Shaban-Nejad, A., 2022. Fine-tuned sentiment analysis of covid-19 vaccine–related social media data: comparative study. J. Med. Internet Res. 24 (10), e40408.
Meng, L., Li, L., Xie, W., Li, Y., Liu, Z., 2023. Time-sequential hesitant fuzzy entropy, cross-entropy and correlation coefficient and their application to decision making. Eng. Appl. Artif. Intel. 123, 106455.
Mu, Z., Zheng, S., & Wang, Q. (2021, August). ACL-RoBERTa-CNN Text Classification Model Combined with Contrastive Learning. In: 2021 International Conference on Big Data Engineering and Education (BDEE) (pp. 193-197). IEEE.
Neviarouskaya, A., Prendinger, H., Ishizuka, M., 2011. SentiFul: a lexicon for sentiment analysis. IEEE Trans. Affect. Comput. 2 (1), 22–36.
Onan, A., 2021. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. Concurrency Computat.: Pract. Exp. 33 (23), e5909.
Park, E., Kang, J., Choi, D., Han, J., 2020. Understanding customers' hotel revisiting behaviour: a sentiment analysis of online feedback reviews. Curr. Issue Tour. 23 (5), 605–611.
Pathak, A.R., Pandey, M., Rautaray, S., 2021. Topic-level sentiment analysis of social media data using deep learning. Appl. Soft Comput. 108, 107440.
Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.
Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
Ptaszynski, M., Rzepka, R., Araki, K., Momouchi, Y., 2014. Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis. Comput. Speech Lang. 28 (1), 38–55.
Puh, K., Bagić Babac, M., 2023. Predicting sentiment and rating of tourist reviews using machine learning. J. Hospitality Tourism Insights 6 (3), 1188–1204.
Qian, C., Mathur, N., Zakaria, N.H., Arora, R., Gupta, V., Ali, M., 2022. Understanding public opinions on social media for financial sentiment analysis using AI-based techniques. Inf. Process. Manag. 59 (6), 103098.
Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
Ren, Z., Shen, Q., Diao, X., Xu, H., 2021. A sentiment-aware deep learning approach for personality detection from text. Inf. Process. Manag. 58 (3), 102532.
Rinaldi, A., Oseguera, O., Tuazon, J., & Cruz, A. C. (2017). End-to-end dialogue with sentiment analysis features. In: HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I 19. Springer International Publishing, pp. 480-487.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.
Sharma, Y., Agrawal, G., Jain, P., & Kumar, T., 2017, December. Vector representation of words for sentiment analysis using GloVe. In: 2017 international conference on intelligent communication and computational techniques (icct). IEEE, pp. 279-284.
Stamatis, C.A., Meyerhoff, J., Liu, T., Sherman, G., Wang, H., Liu, T., Mohr, D.C., 2022. Prospective associations of text-message-based sentiment with symptoms of depression, generalized anxiety, and social anxiety. Depress. Anxiety 39 (12), 794–804.
Suhaimin, M. S. M., Hijazi, M. H. A., Moung, E. G., Nohuddin, P. N. E., Chua, S., & Coenen, F. (2023). Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions. J. King Saud Univ.-Comput. Inform. Sci. 101776.
Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., ... & Wang, H. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137.
Tanna, D., Dudhane, M., Sardar, A., Deshpande, K., Deshmukh, N., 2020, May. Sentiment analysis on social media for emotion classification. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, pp. 911-915.
Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I., 2010, May. Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the international AAAI conference on web and social media, Vol. 4, No. 1, pp. 178-185.
Upadhyaya, A., Fisichella, M., Nejdl, W., 2023. Towards sentiment and temporal aided stance detection of climate change tweets. Inf. Process. Manag. 60 (4), 103325.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inform. Process. Syst. 30.
Wang, Y., Chen, G., Xu, Y.C., Lu, X., 2023. Novel role-wise attention mechanism for predicting purchases made through chat-based online customer services. Decis. Support Syst. 169, 113942.
Wang, B., Shang, L., Lioma, C., Jiang, X., Yang, H., Liu, Q., & Simonsen, J. G. (2020, October). On position embeddings in bert. In: International Conference on Learning Representations.
Wei, Z., Liu, W., Zhu, G., Zhang, S., Hsieh, M.Y., 2022. Sentiment classification of Chinese Weibo based on extended sentiment dictionary and organisational structure of comments. Connect. Sci. 34 (1), 409–428.
Xiao, Z., Nie, Z., Song, C., Chronopoulos, A.T., 2022. An extended attention mechanism for scene text recognition. Expert Syst. Appl. 203, 117377.
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., Le, Q. V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inform. Process. Syst. 32.
Yang, L., Li, Y., Wang, J., Sherratt, R.S., 2020. Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE Access 8, 23522–23530.
Yin, Y., 2020. An emergent algorithmic culture: the data-ization of online fandom in China. Int. J. Cult. Stud. 23 (4), 475–492.
Yu, Y., Wang, Y., Mu, J., Li, W., Jiao, S., Wang, Z., Zhu, Y., 2022. Chinese mineral named entity recognition based on BERT model. Expert Syst. Appl. 206, 117727.
Zhang, L.T., Cassany, D., 2020. Making sense of danmu: coherence in massive anonymous chats on Bilibili. com. Discourse Stud. 22 (4), 483–502.
Zhang, X., Zhou, H., Yu, K., Zhang, X., Wu, X., & Yazidi, A., 2022, June. Sentiment Analysis for Chinese Dataset with Tsetlin Machine. In: 2022 International Symposium on the Tsetlin Machine (ISTM). IEEE, pp. 1-6.
Zhang, M., Liu, Y., 2021. A commentary of TikTok recommendation algorithms in MIT technology review 2021. Fundamental Res. 1 (6), 846–847.
Zhang, L., Pentina, I., 2012. Motivations and usage patterns of Weibo. Cyberpsychol., Behav., Soc. Networking 15 (6), 312–317.
Zhang, T., Yang, K., Alhuzali, H., Liu, B., Ananiadou, S., 2023. PHQ-aware depressive symptoms identification with similarity contrastive learning on social media. Inf. Process. Manag. 60 (5), 103417.
Zhao, N., Jiao, D., Bai, S., Zhu, T., 2016. Evaluating the validity of simplified Chinese version of LIWC in detecting psychological expressions in short texts on social network services. PLoS One 11 (6), e0157947.
Zhao, A., Yu, Y., 2021. Knowledge-enabled BERT for aspect-based sentiment analysis. Knowl.-Based Syst. 227, 107220.
Zuhra, F.T., Saleem, K., 2023. Hybrid embeddings for transition-based dependency parsing of free word order languages. Inf. Process. Manag. 60 (3), 103334.