Journal of Network Intelligence
Taiwan Ubiquitous Information

Volume **9**, Number **1**, February **2024**

Unsupervised Feature Classification-Based Sentiment Analysis of Online Social Network Texts

Xiangzhen Zhou*

Department of Information Engineering
Shengda University
Henan 451191, P. R. China

Faculty Information Science and Technology
National University of Malaysia
Selangor 43600, Malaysia
100618@shengda.edu.cn

Yunfei Zhu

Tsinghua University Press Third Division
Beijing 102611, P. R. China
1468059565@qq.com

*Corresponding author: Xiangzhen Zhou
Received April 23, 2023, revised June 26, 2023, accepted September 29, 2023.

Abstract. *: Comments on Online Social Networks (OSN) contain hidden information about the quality of teaching and learning. However, it takes a lot of time and effort to read these texts item by item. Automating the sentiment analysis of these comments through machine learning techniques can significantly improve the efficiency of text processing. Currently, sentiment analysis techniques for commenting on online social networks have become an important tool for uncovering the deeper psycholinguistic features of university students. The effect of sentiment analysis is not satisfactory due to the unstandardised format of comment information, variable language style and the large amount of noise. To address this problem, an unsupervised class psycholinguistic feature classification method based on sentiment computing is proposed. First, psycholinguistic features are extracted from the comment information through three pre-processing steps: word separation, normalisation and lexical annotation. Then, a 4-dimensional auxiliary sentiment dictionary is combined with a traditional single sentiment polarity value dictionary to compute the sentiment tendencies contained in the comment information from multiple dimensions. Finally, the overall psycholinguistic feature classification is completed by increasing the weight calculation of sentiment words and the calculation of sentiment tendency. test results on both MOOC and Tweet databases show that the proposed unsupervised class classification method is more accurate than other existing dictionary-based unsupervised class classification methods.*
Keywords: Online social networks; Text processing; Sentiment computing; Feature classification; Teaching comments; Sentiment dictionary

1. **Introduction.** The Internet has greatly changed the way people express and communicate with each other. College students can express their opinions and communicate with others through comments on Online Social Networks (OSN). What users say on the Internet is called comment text [1-5]. As we all know, there are many valuable things that can be used in the comment text generated on OSN. For example, people's shopping decisions are increasingly influenced by online comments. For example, if a person wants

to buy an item, he is likely to buy it when the comments he reads are mostly positive. However, if the comments are mostly negative, the user may look for an alternative product. For an individual or an organisation, positive comments represent a financial benefit or reputation. This is the motivation behind the emergence of psycholinguistics. In this work we concentrate on the psycholinguistic features of users' product comments. A similar psycholinguistic phenomenon exists for comment texts on online social networks as for commodity comments.

Sentiment analysis for online social network texts has become a hot research topic [6-9], and sentiment analysis is an important task in the field of natural language processing. Most research has been devoted to extracting summary views from comments using natural language processing techniques and data mining techniques [10-13]. For example, by analysing the textual content of a comment of an electronic product, the user's sentiment tendency towards the price, quality, features and other attributes of the product can be obtained, so that manufacturers can target their products to improve their performance, sellers can always develop more appropriate marketing strategies, and users can choose a more suitable product. Nowadays, almost every website has an area for users to leave comments, and mining this vast amount of views and opinions can provide companies and organisations with information to understand human behaviour, with significant commercial value and social research implications [14-16]. Currently, most unsupervised class sentiment analysis methods rely too much on a single sentiment word, often a pre-determined vocabulary. As a result, they cannot truly reflect the inter-sentence relationships in the text structure.

In order to fully explore the psycholinguistic features of emotional information in university students' comment texts and thus improve the effectiveness of psycholinguistic feature classification, this paper classifies the emotional polarity of comment texts into two categories, positive and negative, and designs and implements an unsupervised class psycholinguistic feature classification method based on emotional computing. The experimental data contains 3 datasets: 1) 1 real data crawled from MOOC; 2) 2 commonly used Tweet datasets. Finally, the performance of the proposed classification method is illustrated through comparative experiments.

1.1. **Related Work.** In general, sentiment analysis techniques can be broadly divided into unsupervised dictionary-based and supervised machine-learning-based approaches. Some researchers have also combined the two approaches for sentiment analysis.

Dictionary-based unsupervised class methods research relies on a sentiment word, often a pre-determined vocabulary. Machine learning-based methods, on the other hand, make use of a mixture of syntactic and linguistic features by means of a multi-medium approach. The sentiment dictionary also plays a major role in this. Shen and Zhang [17] proposed a dictionary of sentiment polarity values and used dictionary-based and custom sentiment score calculation rules to calculate sentiment polarity values. Wang [18] first constructed an emotional corpus by combining the expression pictures and emotional words in Weibo, and optimized the corpus by using the concept of entropy, and trained Bayesian classifier by extracting lexical features to realize emotional classification. Rezwanul et al. [19] use machine learning to select adjectives and verbs in the text as training features, and use feature dimension reduction method based on hierarchical structure. The characteristic polarity value is calculated by symbolic expression, and the weight is calculated by polarity value. Finally, SVM and KNN classifier are used to classify the text into three types: positive, negative and neutral. However, supervised machine learning-based methods require a large amount of annotated data during the training phase, and therefore have

a high computational overhead in terms of CPU processing, memory requirements and training time. In addition, supervised classifiers are difficult to extend to other domains.

1.2. **Motivation and contribution.** Therefore, an unsupervised class psycholinguistic feature classification method based on multidimensional sentiment computation is proposed to address the problems that arise above. First, word-level, phrase-level and sentence-level psycholinguistic features are extracted through a pre-processing step. Then, a 4-dimensional auxiliary sentiment dictionary is combined with a traditional single sentiment polarity value dictionary to compute the sentiment tendency contained in the comment message from multiple dimensions. Finally, the overall feature classification is completed by adding weight calculations for sentiment words and sentiment tendency calculations.

The main innovations and contributions of this study are shown below:

1) The extraction of word multi-level psycholinguistic features from comment information through three pre-processing steps of word separation, normalisation and lexical annotation, enabling the identification of all types of characters, e.g. words, links, numbers and time, and in particular various forms of epigraphic symbols and their corresponding emotional polarity.

2) A 4-dimensional auxiliary sentiment dictionary is combined for sentiment analysis, thus reflecting more realistically the inter-sentence relationships in the text structure and integrating deep syntactic features.

2. Data pre-processor.

2.1. **Problem description.** Considering that classifiers using machine learning are usually more difficult to interpret and therefore not conducive to modification, generalisation or extension, this paper chooses an unsupervised class approach based on a dictionary of sentiment polarity values.

However, the main problems with existing research are as follows: 1) existing research has mainly focused on normal canonical written texts, but comment messages may also contain many abbreviations, symbolic expressions, topic tags, slang, link addresses, etc., which makes psycholinguistic feature extraction difficult, as confirmed in Dolynska et al. [20]; 2) the performance of traditional single sentiment polarity-valued dictionaries do not perform well enough to truly reflect inter-sentence relations in text structure. Yuan et al. [21] attempted to address this problem by combining the weight calculation of sentiment words, but still failed to reflect deep syntactic features.

Comment messages on online social networks are text messages with informal spelling and sentence structure, often containing a large number of misspellings, random spaces, punctuation etc. This makes existing low-level text pre-processing tasks such as word separation and normalisation inefficient. We have therefore designed an optimised preprocessor to address the problem of informal writing in social media text messages. Some of the outputs of the data preprocessor are shown in Table 1.

Table 1. Selected output examples from the data pre-processor.

	0	**1**	**2**	**3**	**4**	**5**
token	haha	^_^	!!!	LOOOOL	a	🙂
t_norm	#goodday	[E1+]	!	LOL	A	[E2+]
t_tag	#	.	?	.	~	o(>_<)o

As shown in Table 1, some of the output examples include normalised characters, character tags and lexical annotations. Character tags are assigned by the lexer and include letters, slang, punctuation, symbolic expressions, topic tags, link addresses, etc.

2.2. **Splitting words.** Given a piece of text, the task of splitting it into separate small parts, called characters, which can be a word or other meaningful semantic units, such as numbers, usernames, topic tags, email addresses, symbolic emojis, etc.

Considering that people type mostly randomly nowadays, with a variety of incorrect writing formats, we cannot simply separate text by spaces. We used the open-source FreelCTCLAS Tokenizer from CAS [22] to address most common stylistic errors in social media texts, such as repeated punctuation, extra spaces before tag symbols or missing spaces after punctuation.

The FreelCTCLAS Tokenizer process matches different types of characters by means of regular expressions, which are immediately labelled accordingly. Care must be taken here to exclude characters that are already labelled before proceeding with the match. For example, symbolic expressions and numbers are predominantly labelled before punctuation. Symbolic expressions usually contain punctuation marks. The decimal point in a number is also punctuated. Similarly, commas are used in larger numbers to improve readability.

2.3. **Standardisation.** As non-standard characters are common in social media messages. For example, many people will use repeated letters or symbolic emojis to emphasise a sentiment, emotion or point of view, so a 4-step process is used to standardise the process:

(1) Repeat letter writing.

Repetitive letter writing is the use of repeated letters in text to increase the strength of expression of words. For example, the phonetic code for both lol and lloooollll is L400. to solve the problem of using repeated letters, we use the phonetic code to find the index of each word and identify matching options, and calculate the Levenshtein distance between the input and each matching option to measure its similarity to return the best match.

(2) Symbolic expressions.

Use regular expressions to identify the polarity of a facial expression symbolic expression. For example, if we use [E1+] for positive symbolic expressions (ˆ_ˆ), we use [E1-] for negative symbolic expressions.

(3) Picture expressions.

Similar to symbolic emoticons, users often use picture emoticons to convey positive or negative emotions and feelings. We have standardised all picture emojis into predefined characters, such as [E2+] and [E2-]. These three predefined characters indicate positive and negative respectively.

4) Text expressions

Similar to writing a letter repeatedly, it is common for users to expand these characters for emphasis, e.g. hhahahahahah. We find these text expressions by matching regular expressions containing at least k repeated letters (k=2), and then normalizing each text expression to its core form. For example, hhahahahahah becomes haha.

2.4. **Lexical annotation.** The FreelCTCLAS Tokenizer used has a self-contained annotator. However, in order to further improve the reliability of lexical annotation in the presence of spelling errors, we have optimised the splitter.

Firstly, the original message is split into characters and a new message is generated. This new message has two changes compared to the original message: 1) it is corrected for repeated letters and changed to the standard form, and 2) missing spaces are automatically

added and redundant spaces are removed. This ensures that the lexically annotated separator gets the same character set as the optimised separator.

2.5. **Psycholinguistic feature extraction.** After the three pre-processing steps described above, psycholinguistic features at word level, phrase level and sentence level were extracted in order to be effective. We divided the feature extraction and sentiment calculation into two separate parts.

There are certain differences in the psycholinguistic features of positive and negative comments. First, there are more negative words in negative comments than in positive comments. In addition, there are more verbs than nouns in negative comments as opposed to positive comments. Secondly, the frequency patterns of pronouns in negative and positive comments do not coincide. In particular, first-person singular pronouns occur more frequently in negative comments than in positive comments. Based on this preliminary observation, we devised a set of features to represent the psycholinguistic characteristics of the comments.

$$T_i^{\text{new}} = T_i^{\text{old}} \cdot 2^{\log_{10}\left(\sum_p p \in P\right)} \tag{1}$$

where T_i^{new} is the set of feature items for character i after the update, T_i^{old} is the set of feature items for character i before the update, p is the set of valid words that can be used for sentiment analysis, and P is the set of negation words and pronouns.

For word-level, phrase-level and sentence-level comments, psycholinguistic features are extracted using the feature extraction method shown in Equation (1), which allows for the identification of all types of characters, especially the various forms of epigraphs and their corresponding emotional polarity.

3. Classification of unsupervised class psycholinguistic features.

3.1. **Feature classification framework.** To implement unsupervised class psycholinguistic feature classification, an affective computational classification framework consisting of three separate components together was designed, as shown in Figure 1.

3.2. **Linguistic feature extraction.** We find that users are usually careless when using this way to emphasize, especially those who use touch-screen mobile phones often forget to capitalize the first or last letter.

In order to avoid missing these situations, if 75% of the letters in a character are capitalized, we consider that the character is all capitalized. If a character i is capitalized, $s_i^{ALLCaps} = 1$, otherwise $s_i^{ALLCaps} = 0$.

$$T_i^{\text{new}} = T_i^{\text{old}} \cdot 2^{\log_{10}\left(1+s_t^{ALLCaps}\right)} \tag{2}$$

If the character i is not capitalized, $T_i^{\text{new}} = T_i^{\text{old}}$, otherwise $T_i^{\text{new}} = 1.232 \times T_i^{\text{old}}$.

Because users will emphasize their feelings or opinions by reusing certain letters or symbols, we assign an elongation factor $\frac{s_i^{ExLen}}{i}$ to each character i. This elongation factor can reflect the different degrees of the original word length and the standardized word length. Considering that letter elongation can also enhance the emotional strength of characters, we can update T_i^{new} as follows.

$$T_i^{\text{new}} = T_i^{\text{old}} \cdot 2^{\log_{10}\left(s_t^{ExLens}\right)} \tag{3}$$

where s_i^{ExLen} is the ratio between the original character and the normalized character.

Reinforcing words (such as *very*, *really*, *extreme*) and weakening words (such as *hardly*, *barely*, *a little*), all belong to modifiers, which have little practical significance in themselves but can enhance or weaken the emotional strength of the modified words. Similar

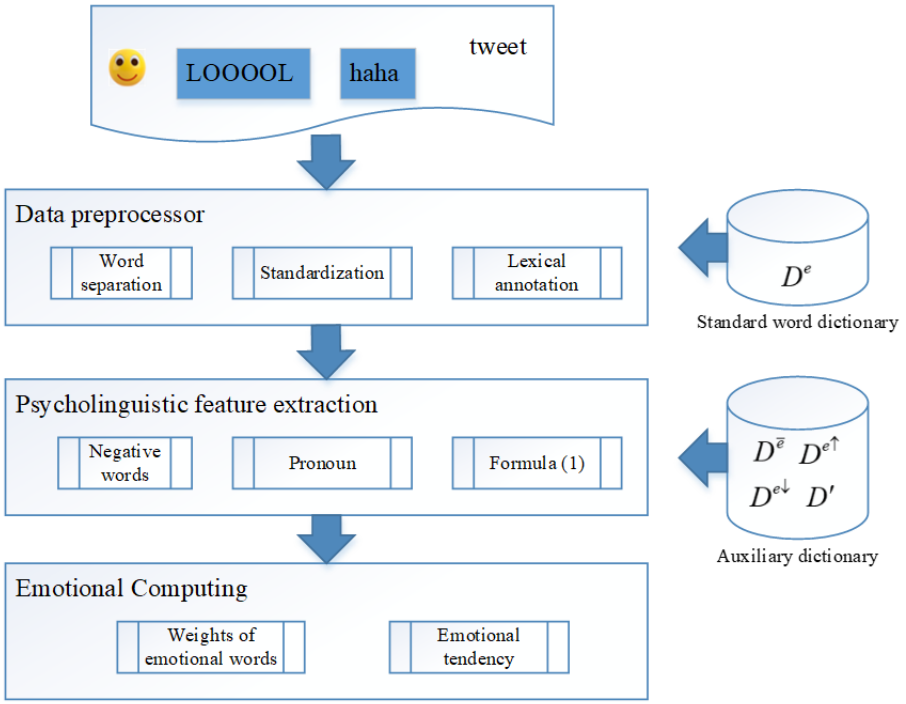


Figure 1. Unsupervised class psycholinguistic feature classification framework

to matching adjective noun pairs, we still use regular expressions to match part-of-speech tags, and at the same time use modifier dictionaries $D^{e\uparrow}$ and $D^{e\downarrow}$ to find out all modifiers and modified words. We will mark a series of modifiers of the character i with M_t^{DM}.

$$s_i^{DM} = \sum_{m \in M_i^{DM}} \left(1 + s_m^{AllCaps} + s_m^{ExLen}\right) \tag{4}$$

where s_t^{DM} is the elongation factor of the modifier i, m is a modifier and M^{DM} is the set of modifiers.

Accordingly, we need to update the influence factor T_i^{new} of emotional score.

$$T_i^{\text{new}} = T_i^{\text{old}} \cdot 2^{\log_{10}\left(s_t^{DM}\right)} \tag{5}$$

For negative words, unsupervised methods usually change the polarity of negative words in emotional dictionaries, while supervised methods usually change negative words in the training stage, so that the original text can be distinguished from the text after using negative words. In order to identify negative words, we need a negation dictionary $D^{\bar{e}}$. Update the influence factor T_i^{new} of the character i modified by negative words by the following formula.

$$T_i^{\text{new}} = T_i^{\text{old}} \cdot (-1) \tag{6}$$

3.3. **A multidimensional dictionary of affective polarity values.** We use a sentiment polarity dictionary and some auxiliary dictionaries to implement psycholinguistic feature analysis of social media texts. The sentiment polarity dictionary provides sentiment polarity values for characters that are likely to have sentiment polarity, while the auxiliary dictionaries are used for better data pre-processing and structured feature extraction.

A dictionary of affective polarity values is a dictionary containing a large number of words with affective polarity and information about their corresponding affective polarity.

There has been a great deal of work devoted to building dictionaries of sentiment polarity values. Sentiment polarity dictionaries vary in the words they contain, for example some dictionaries do not contain swear words such as WTF or internet slang abbreviations such as LOL, but such words are essential in the work with social media messages.

There are also different types of affective polarity in dictionaries, some of which have specific scores, while others may only give positive or negative labels. For example, VADER (Valence Aware Dictionary for sEntiment Reasoning), created by Hutto et al. [23] in 2014, is one of the latest freely available dictionaries of sentiment polarity values, specifically for OSN text messages such as Tweets.

In addition to the standard word dictionary in sentiment dictionary, we make use of a number of auxiliary dictionarys to calculate the sentiment tendencies contained in the comment information from four dimensions (negation dictionary $D^{\bar{e}}$, enhanced dictionary of modifiers $D^{e\uparrow}$, diminished modifier dictionary $D^{e\downarrow}$, dictionary of common slang D'). The four auxiliary dictionarys are mainly used in the normalisation process as well as in the feature extraction process. The definitions of the four auxiliary dictionarys are given in Table 2.

Table 2. Definitions of the 4 supporting dictionaries

Symbols	Definition	Example
D^e	Standard Word Dictionary	happy, move, good
$D^{\bar{e}}$	Negation Dictionary	not, neither, never
$D^{e\uparrow}$	Enhanced Dictionary of Modifiers	very, really, extremely
$D^{e\downarrow}$	Diminished Modifier Dictionary	hardly, slightly, a little
D'	Dictionary of Common Slang	lol: laugh out loud

The Negation Dictionary $D^{\bar{e}}$ is a subset of the standard word dictionary D^e created by hand. $D^{\bar{e}}$ mainly includes words and phrases with a negative effect, such as don't, haven't, etc. The enhanced modifier dictionary $D^{e\uparrow}$ and the diminished modifier dictionary $D^{e\downarrow}$ were also created manually. The combination of the two can influence the sentiment of a comment by affecting the sentiment intensity of the sentence in which it is placed.

3.4. **Basic Emotional Score Calculation.** The emotion calculator uses a plurality of emotion dictionaries and the output results of the feature extraction part to assign an emotion score to each character. Let L be the dictionary set of emotional polarity values that we choose. Through these dictionaries, we can get the basic emotional polarity score of each character.

$$s_i = \begin{cases} \frac{\sum\limits_{l \in L_i} \text{score}(l,i)}{|L_i|}, L_i \neq 0 \\ 0, L_i = 0 \end{cases} \tag{7}$$

where $L_i = \{l \in L | i \in l\}$ is a subset of the emotional polarity dictionaries containing the character i, $|L_i|$ represents the number of emotional polarity value dictionaries, and score is the basic emotional polarity value of each independent text character i given in the dictionary l.

Update the emotional scores one by one according to linguistic features. The basic emotional score of each character i is updated by using the emotional score influence factor T_i.

$$s_i^{\text{new}} = s_i^{\text{old}} \cdot T_i \tag{8}$$

Then, using the extracted independent linguistic feature information. The referenced single character is updated as follows.

$$s_i^{\text{new}} = (-1) \cdot s_i^{\text{old}} \tag{9}$$

3.5. **Psycholinguistic feature classification based on affective computing.** We complete the overall psycholinguistic feature classification by adding a weighting calculation for the emotion words and an emotion propensity calculation.

Weighting of sentiment words: $S = \{s_1, s_2, ..., s_v\}$ is used to denote sentiment words and V is the number of sentiment words. After pre-processing, the comment text will consist of some words or words with meaning. Let the input comment text be M, and use FreeLCTCLAS Tokenizer to classify M into words. Suppose $m_i \in M$, then its feature vector is represented as

$$\vec{f} = \begin{bmatrix} f_{i1} \\ \dots \\ f_{iv} \end{bmatrix} \tag{10}$$

where f_{ij} indicates the importance of the sentiment word s_j in the comment m_i. The greater the value of f_{ij}, the greater the contribution of s_j to emotional judgment of m_i.

The traditional TF-IDF [24] was used to represent the weights of sentiment words.

$$\text{score}(s, \bar{m}) = \frac{tf(s, \bar{m}) \times \log\left(N/n_s + 0.01\right)}{\sqrt{\sum\limits_{s \in \bar{m}} \left[f(s, \bar{m}) \times \log\left(N/n_s + 0.01\right)\right]^2}} \tag{11}$$

where $\text{score}(s, \overline{m})$ denotes the weight of sentiment word s in the comment $\overline{m}$, $tf(s, \overline{m})$ denotes the word frequency of sentiment word s in the comment $\overline{m}$, N is the number of training samples, and n_s is the number of documents in which sentiment word s occurs in the training sample. The denominator is the normalization factor such that the weight of each feature word is between [0, 1].

Sentiment propensity calculation: For each sentiment word $s_j \in S$, assume that its polarity value is $\omega(s_j)$. If $\omega(s_j) > 0$, then s_j is a positive sentiment word, and if $\omega(s_j) < 0$, then s_j is a negative sentiment word. $|\omega(s_j)|$ denotes the sentiment intensity of s_j. Assuming $\omega(s_j) = \omega_j$, the vector of sentiment word polarity values is as follow:

$$\vec{\omega} = \begin{bmatrix} \omega_1 \\ \dots \\ \omega_v \end{bmatrix} \tag{12}$$

For each comment m_i , the polarity value of the comment is as follow:

$$\alpha\left(m_i\right) = \text{cosine}\left(\overrightarrow{f_i}, \vec{\omega}\right) = \frac{\overrightarrow{f_i}^T \cdot \vec{\omega}}{|\vec{\omega}|} \tag{13}$$

Assuming $\alpha(m_i) = \alpha_i$, α_i is the polarity value of the comment m_i. If $\alpha_i > 0$, the overall psycholinguistic profile is judged to be in the positive category, and if $\alpha_i < 0$, the overall psycholinguistic profile is judged to be in the negative category.

4. **Experimental results and analysis.**

4.1. **Experimental data.** The experimental data contains 3 datasets: 1) real data crawled from the MOOC online learning platform (www.icourse163.org/); 2) VADER-Tweet dataset [25]; 3) SemEval-2013 dataset [26].

The MOOC dataset contains 100 topics, with a total of 11 646 comments, with 8,998 sample data in the positive category and 2,648 sample data in the negative category. For all data, five people will mark the positive and negative of each Weibo, and if the marking results are inconsistent, they will negotiate and then re-mark. The VADER-Tweet dataset was manually annotated with a large number of Tweets and noise filtered by using constraints, thus ensuring the quality of annotation of Tweet sentiment polarity as much as possible. SemEval-2013 is a dataset specifically designed to test the sentiment analysis of Twitter texts, containing 1028 positive and 2628 negative texts, for a total of 3656 texts. Three experimental data sets are shown in Table 3.

Table 3. Experimental data sets

Data sets	Tags	Number
MOOC (11646)	Positive category samples	8998
	Negative category samples	2648
VADER-Tweet (4200)	Positive category samples	2897
	Negative category samples	1303
SemEval-2013 (3656)	Positive category samples	1028
	Negative category samples	2628

4.2. **Assessment indicators.** The results of the psycholinguistic feature classification were evaluated using three metrics, namely Precision, Recall and F-Measure, with the total number of documents in a collection of documents being D. The text representation is shown in Table 4. Equation (14) is the calculation of the accuracy P (Precision);

Table 4. Text representation

	A sample of the actual negative category	**A sample of what is not actually a negative category**
Sample marked as negative category	a	b
Samples marked not in the negative category	c	d

Equation (15) is the calculation of the recall R (Recall); Equation (16) is the calculation of the comprehensive evaluation index F (F-Measure) of classification performance.

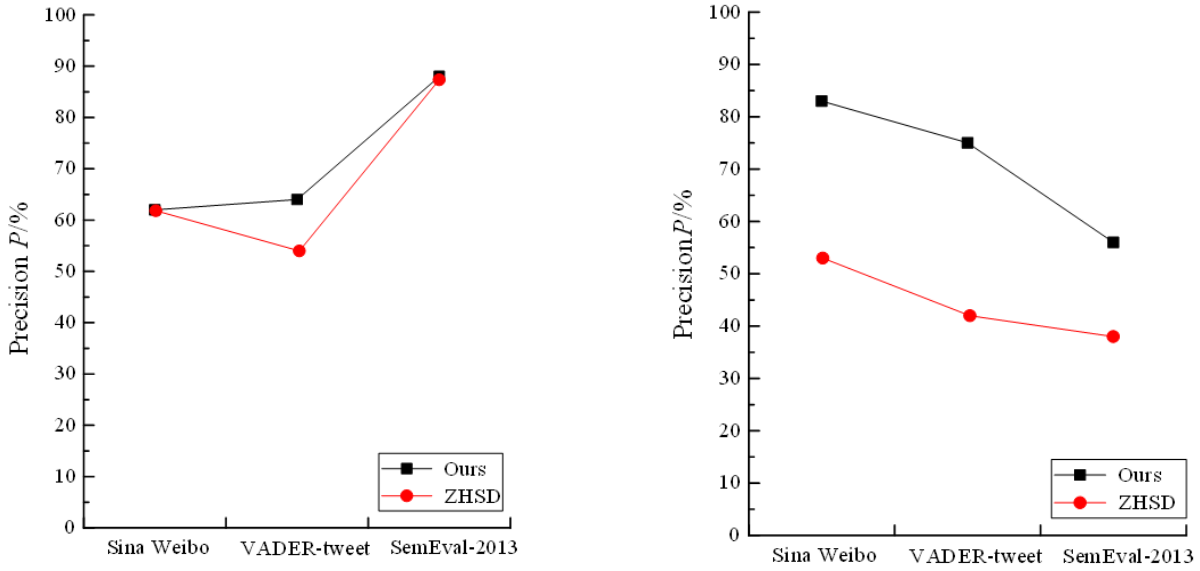
$$P = \frac{a}{a+b} \times 100\% \tag{14}$$

$$R = \frac{a}{a+c} \times 100\% \tag{15}$$

$$F = \frac{2 \times P \times R}{P+R} \tag{16}$$

4.3. **Performance comparison with a single sentiment polarity-valued dictionary.** In order to illustrate the performance of the proposed method for the classification of psycholinguistic features, two sets of comparison experiments were experimented on each of the three datasets, and the results are shown in Figure 2 and Figure 3.

ZHSD is the "Word Collection for Sentiment Analysis (beta version)" published by China Knowledge Network [27]. DLSD is a Chinese DLSD is a Chinese dictionary of emotion provided by Dalian University of Technology [28].

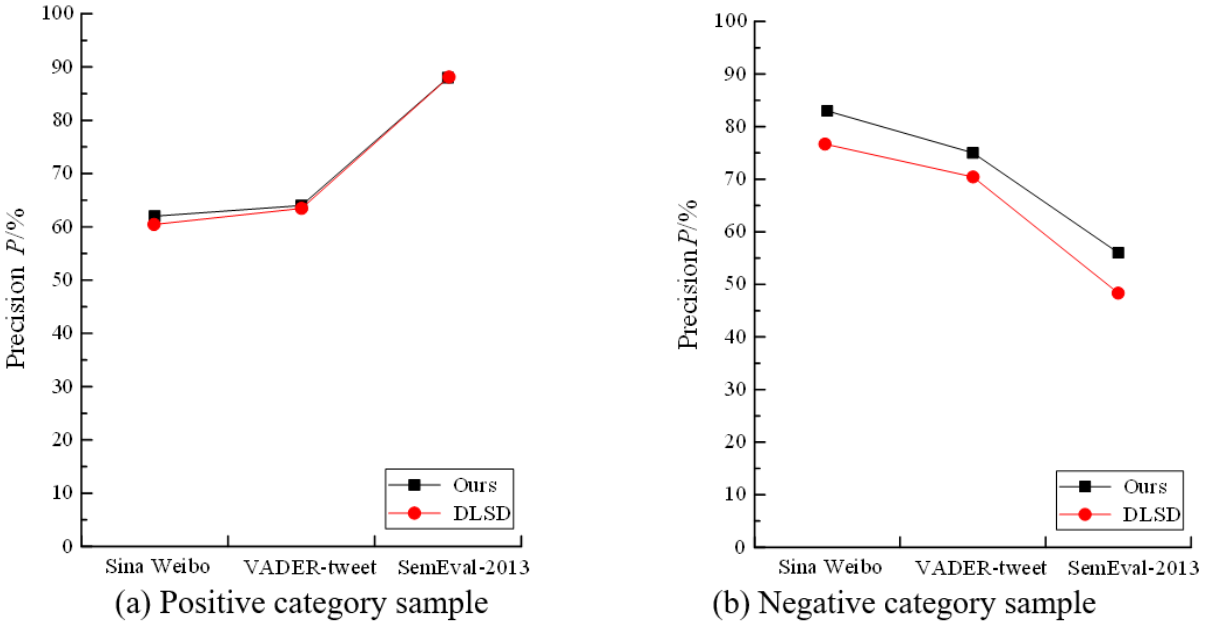


(a) Positive category sample (b) Negative category sample

Figure 2. Performance comparison with ZHSD dictionary

As shown in Figure 2(a), for the positive class samples, the classification performance of the proposed method is significantly better than that of the single dictionary ZHSD on the VADER-Tweet dataset. However, the classification performance on the remaining two datasets is almost identical. As shown in Figure 2(b), for the negative class samples, it can be seen that the classification performance of the proposed method is significantly better than that of the single dictionary ZHSD on all three datasets. As shown in Figure 3(a), for the positive class samples, the difference between the classification performance of the proposed method on the VADER-Tweet dataset and the DLSD dictionary is small. As shown in Figure 3(b), for the negative class samples, it can be seen that the classification performance of the proposed method is slightly better than that of the single dictionary DLSD on all three datasets.

Overall, unlike existing unsupervised classes of single sentiment polarity-valued dictionaries, the proposed method uses a 4-dimensional auxiliary sentiment dictionary, which allows the calculation of the sentiment tendency contained in a comment message from multiple dimensions. Although the single dictionary also contains a large number of other non-standard characters such as common symbolic expressions and slang abbreviations, the data preprocessor used determines the sentiment polarity of the comments directly based on the sentiment polarity of symbolic expressions and picture expressions, and no longer detects sentiment polarity characters in the text, which helps to improve the accuracy of the final classification results to some extent.



(a) Positive category sample (b) Negative category sample

Figure 3. Performance comparison with DLSD dictionary

4.4. **Performance comparison with a composite sentiment polarity-valued dictionary.** In addition to the unsupervised class singleton sentiment polarity value dictionary, the proposed method was compared with an unsupervised class method based on a compound sentiment polarity value dictionary.

Li et al. [29] constructed a compound polarity dictionary that included a base dictionary, a domain dictionary, a network word dictionary, and a modifier dictionary. Then, sentiment analysis was performed based on the polarity dictionary, including calculating the polarity of polar phrases and calculating the polarity of sentences and the whole. The proposed method was compared with the method based on a composite sentiment polarity value dictionary, as shown in Tables 5, 6 and 7.

Table 5. Accuracy P comparison results

Data sets		Dictionary of compound emotional polarity values %	Methodology proposed / %
MOOC	Positive category samples	0.60	0.62
	Negative category samples	0.71	0.83
VADER-Tweet	Positive category samples	0.63	0.64
	Negative category samples	0.67	0.75
SemEval-2013	Positive category samples	0.88	0.88
	Negative category samples	0.51	0.56

As can be seen from Tables 5, 6 and 7, for the negative class samples in the three datasets, the proposed method has higher accuracy P, recall R, and F-value than the classification method based on the composite polarity dictionary. For the positive class samples, there is little difference between the performance of the proposed method and the classification method based on the composite sentiment polarity value dictionary. The reason for this is that the proposed method takes into account not only the word frequency information and sentiment information of the sentiment words themselves, but also the intensity of the sentiment expressed by the sentiment words. We use not only

Table 6. Comparison results for recall R

Data sets		Dictionary of compound emotional polarity values %	Methodology proposed / %
MOOC	Positive category samples	0.67	0.68
	Negative category samples	0.59	0.65
VADER-Tweet	Positive category samples	0.68	0.67
	Negative category samples	0.45	0.53
SemEval-2013	Positive category samples	0.86	0.87
	Negative category samples	0.42	0.56

Table 7. Comparative results of the integrated evaluation indicators F

Data sets		Dictionary of compound emotional polarity values %	Methodology proposed / %
MOOC	Positive category samples	0.61	0.62
	Negative category samples	0.60	0.72
VADER-Tweet	Positive category samples	0.63	0.63
	Negative category samples	0.54	0.69
SemEval-2013	Positive category samples	0.83	0.82
	Negative category samples	0.48	0.61

the sentiment intensity of the sentiment word, but also the weight of the sentiment word, which means that the sentiment propensity is calculated differently. It can be seen that the data pre-processing, as well as the psycholinguistic features that are fully utilised, can significantly improve the final classification accuracy.

4.5. **Discussion.** From the above experimental results, it can be seen that the proposed method has a higher classification accuracy than the classification methods based on single or compound sentiment polarity value dictionaries. The drawback of this study is that the classification results are better for the negative category samples, but the improvement in classification accuracy for the positive category samples is not satisfactory. In other words, the proposed method is easier to analyse the negative comments. One of the reasons for this is that for comment information containing both positive and negative words, and the negative words have a somewhat larger score and therefore determine the final polarity.

5. **Conclusion.** This paper relies on a sentiment polarity value dictionary and four auxiliary dictionarys to implement psycholinguistic feature analysis of comment texts on OSN. The psycholinguistic features are extracted from the comment information through three pre-processing steps: word division, normalisation and lexical annotation, and the overall psycholinguistic feature classification is completed by adding the weight calculation of sentiment words and the calculation of sentiment propensity. Three datasets were used to validate the efficiency of the proposed method. The test results show that the proposed method has a higher classification accuracy than the classification methods based on single or compound sentiment polarity value dictionaries. The shortcoming of this study is that the classification results are better for the negative category samples, but the improvement in classification accuracy for the positive category samples is not satisfactory. In other words, the proposed method is easier to analyse the negative comments. One of the reasons for this is that targeting comment information that contains both positive and negative words, with the negative words having a somewhat larger score, determines

the final polarity. Further research will be conducted on how to solve the problem of low classification accuracy of positive samples.

Acknowledgment. This work supported by the project "2022 Science and Technology Research Project of the Department of Science and Technology of Henan Province, NO.222102210290", and "School level 2021 Applied Basic Research and Application Special Project, NO.SD-ZDIAN2021-05".

REFERENCES

[1] T.-Y. Wu, X. Guo, L. Yang, Q. Meng, and C.-M. Chen, "A lightweight authenticated key agreement protocol using fog nodes in Social Internet of vehicles," *Mobile Information Systems*, vol. 2021, Article ID 3277113, 2021.

[2] T.-Y. Wu, J. C.-W. Lin, U. Yun, C.-H. Chen, G. Srivastava, and X. Lv, "An efficient algorithm for fuzzy frequent itemset mining," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5787-5797, 2020.

[3] T.-Y. Wu, J. Lin, Y. Zhang, and C.-H. Chen, "A Grid-Based Swarm Intelligence Algorithm for Privacy-Preserving Data Mining," *Applied Sciences*, vol. 9, no. 4, Article ID 774, 2019.

[4] C.-M. Chen, Z. Li, S. Kumari, G. Srivastava, K. Lakshmanna, and T. R. Gadekallu, "A provably secure key transfer protocol for the fog-enabled Social Internet of Vehicles based on a confidential computing environment," *Vehicular Communications*, vol. 39, Article ID 100567, 2023.

[5] Z. Li, Q. Miao, S. A. Chaudhry, and C.-M. Chen, "A provably secure and lightweight mutual authentication protocol in fog-enabled social Internet of vehicles," *International Journal of Distributed Sensor Networks*, vol. 18, no. 6, Article ID 155013292211043, 2022.

[6] E. K. Wang, C.-M. Chen, S. M. Yiu, M. M. Hassan, M. Alrubaian, and G. Fortino, "Incentive evolutionary game model for opportunistic social networks," *Future Generation Computer Systems*, vol. 102, pp. 14-29, 2020.

[7] H. E. Sama, "Naïve Bayes Twitter Sentiment Analysis In Visualizing The Reputation Of Communication Service Providers: During Covid-19 Pandemic," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 5, pp. 1753–1764, 2021.

[8] W. Zhao, Z. Guan, L. Chen, and X. He, "Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 185–197, 2018.

[9] P. Liu, L. Zhang, and J. A. Gulla, "Multilingual Review-aware Deep Recommender System via Aspect-based Sentiment Analysis," *ACM Transactions on Information Systems*, vol. 39, no. 2, pp. 1–33, 2021.

[10] S. Yadav, R. S. Suhag, and K. V. Sriram, "Stock price forecasting and news sentiment analysis model using artificial neural network," *International Journal of Business Intelligence and Data Mining*, vol. 19, no. 1, pp. 113–121, 2021.

[11] G. Bologna and Y. Hayashi, "A Rule Extraction Study from SVM on Sentiment Analysis," *Big Data and Cognitive Computing*, vol. 2, no. 1, pp. 6–14, 2018.

[12] "USD/PKR Exchange Rate Forecasting Using Sentiment Analysis of Twitter Data," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 3451–3461, 2021.

[13] M. Ghiassi and S. Lee, "A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach," *Expert Systems with Applications*, vol. 106, pp. 197–216, 2018.

[14] M. Hunold, R. Kesler, and U. Laitenberger, "Rankings of Online Travel Agents, Channel Pricing, and Consumer Protection," *Marketing Science*, vol. 39, no. 1, pp. 92–116, 2020.

[15] A. M. Abubakar and M. Ilkan, "Impact of online WOM on destination trust and intention to travel: A medical tourism perspective," *Journal of Destination Marketing & Management*, vol. 5, no. 3, pp. 192–201, 2016.

[16] R. Nie, Z. Tian, J. Wang, and K. S. Chin, "Hotel selection driven by online textual reviews: Applying a semantic partitioned sentiment dictionary and evidence theory," *International Journal of Hospitality Management*, vol. 88, Article ID 102495, 2020.

[17] W. Shen and S. Zhang, "Emotional Tendency Dictionary Construction for College Teaching Evaluation," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 13, no. 11, pp. 117–126, 2018.

[18] H. Wang, "An Unsupervised Microblog Emotion Dictionary Construction Method and Its Application on Sentiment Analysis," *Journal of Information and Computational Science*, vol. 12, no. 7, pp. 2729–2739, 2015.

[19] M. Rezwanul, A. Ali, and A. Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 19–21, 2017.

[20] L. Dolynska, Y. Naumova, and N. Shevchenko, "Psycholinguistic Features of Students' Acquisition of Visual-Semantic Image of a Hieroglyph in Studying Japanese," *Psiholìngvìstika*, vol. 27, no. 1, pp. 30–51, 2020.

[21] L. Yuan, D. Li, S. Wei, and M. Wang, "Research of Deceptive Review Detection Based on Target Product Identification and Metapath Feature Weight Calculation," *Complexity*, vol. 2018, pp. 1–12, 2018.

[22] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[23] W. Liu, G. Cao, and J. Yin, "Bi-Level Attention Model for Sentiment Analysis of Short Texts," *IEEE Access*, vol. 7, pp. 119813–119822, 2019.

[24] U. Rani and K. Bidhan, "Comparative Assessment of Extractive Summarization: TextRank, TF-IDF and LDA," *Journal of Scientific Research*, vol. 65, no. 01, pp. 304–311, 2021.

[25] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, "Tweets Classification on the Base of Sentiments for US Airline Companies," *Entropy*, vol. 21, no. 11, pp. 1078–1090, 2019.

[26] Sefa Sahin Koc, M. Ozer, Ismail Hakki Toroslu, Hasan Davulcu, and J. D. Jordan, "Triadic co-clustering of users, issues and sentiments in political tweets," *Expert Systems with Applications*, vol. 100, no. 15, pp. 79–94, 2018.

[27] W. Ansar, S. Goswami, A. Chakrabarti, and B. Chakraborty, "An efficient methodology for aspect-based sentiment analysis using BERT through refined aspect extraction," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 5, pp. 9627–9644, 2021.

[28] I. Gupta and N. Joshi, "Feature-Based Twitter Sentiment Analysis with Improved Negation Handling," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 917–927, 2021.

[29] J. Li and P. Meesad, "Combining Sentiment Analysis with Socialization Bias in Social Networks for Stock Market Trend Prediction," *International Journal of Computational Intelligence and Applications*, vol. 15, no. 1, 1650003, 2016.