BMC Psychology

RESEARCH

Open Access

Evaluation of emotion classification schemes in social media text: an annotation-based approach

Fa Zhang[1*], Jian Chen[1], Qian Tang[1] and Yan Tian[1]

Abstract

Background Emotion analysis of social media texts is an innovative method for gaining insight into the mental state of the public and understanding social phenomena. However, emotion is a complex psychological phenomenon, and there are various emotion classification schemes. Which one is suitable for textual emotion analysis?

Methods We proposed a framework for evaluating emotion classification schemes based on manual annotation experiments. Considering both the quality and efficiency of emotion analysis, we identified five criteria, which are solidity, coverage, agreement, compactness, and distinction. Qualitative and quantitative factors were synthesized using the AHP, where quantitative metrics were derived from annotation experiments. Applying this framework, 2848 Sina Weibo posts related to public events were used to evaluate the five emotion schemes: SemEval's four emotions, Ekman's six basic emotions, ancient China's Seven Emotions, Plutchik's eight primary emotions, and GoEmotions' 27 emotions.

Results The AHP evaluation result shows that Ekman's scheme had the highest score. The multi-dimensional scaling (MDS) analysis shows that Ekman, Plutchik, and the Seven Emotions are relatively similar. We analyzed Ekman's six basic emotions in relation to the emotion categories of the other schemes. The correspondence analysis shows that the Seven Emotions' joy aligns with Ekman's happiness, love demonstrates a significant correlation with happiness, but desire is not significantly correlated with any emotion. Compared to Ekman, Plutchik has two more positive emotions: trust and anticipation. Trust is somewhat associated with happiness, but anticipation is weakly associated with happiness. Each emotion of Ekman's corresponds to several similar emotions in GoEmotions. However, some emotions in GoEmotions are not clearly related to Ekman's, such as approval, love, pride, amusement, etc.

Conclusion Ekman's scheme performs best under the evaluation framework. However, it lacks sufficient positive emotion categories for the corpus.

Keywords Emotion classification scheme, Evaluation, Social media, Annotation

*Correspondence:
Fa Zhang
richter2000@163.com
[1]Beijing Institute of Technology, Zhuhai School, No.6, JinFeng Road, Zhuhai, Guangdong Province 519088, China

Background

In the age of Web 2.0, many people use online social media. Social media reflects the emotions, attitudes, and opinions of Internet users. Sentiment analysis, the basic task of which is to determine the polarity of a text, such as positive, negative, or neutral [1], has been widely used in social media [2]. Beyond polarity, emotion analysis could identify the types of emotions such as joy, anger, sadness, and fear, helping to understand the mental state more accurately. Emotion analysis is an important topic that has received wide attention [3]. Emotion analysis of social media also has a wide range of applications [4]. For instance, during the COVID-19 epidemic, emotion analysis was used to understand people's emotions and assess policy effects [5]. Some companies conduct emotion analysis on online reviews to understand the user experience and to enhance product development [6].

A fundamental part of emotion analysis is the selection of an emotion model. An emotion model is a theoretical framework for describing, explaining, or predicting human emotions and affective processes. There are many emotion models, roughly categorized as discrete and dimensional [7]. The discrete approach suggests that humans have discrete, distinguishable emotions. A component of discrete emotion models is how to categorize emotions, which in this paper is called an emotion classification scheme. If a discrete emotion approach is adopted, emotion analysis is a multi-classification problem [4]. Lexicon-based methods and machine learning methods are commonly used for emotion classification. The lexicon-based method depends mainly on the lexicon and rules. The machine learning method can obtain better classification results, but it needs to be trained with a large corpus. Regardless of which method is used, the problem of choosing an emotion scheme is faced. There are various emotion classification schemes. Which one is appropriate for emotion analysis?

A systematic approach is required for selecting an emotion classification scheme. The emotion schemes have complex effects on the quality and efficiency of emotion analysis. For example, a suitable scheme can enhance the performance of machine learning models and achieve better application results [8]. In supervised learning, annotated datasets are required, and the emotion scheme has an impact on annotation efficiency.

In this paper, we propose an Analytic Hierarchy Process (AHP) evaluation framework based on annotation experiments. This framework used five criteria, which are solidity, coverage, agreement, compactness, and distinction, to evaluate the emotion schemes. This framework could combine qualitative and quantitative factors. Applying this framework, we collected Sina Weibo posts related to public events, conducted annotation experiments, and evaluated five emotion classification schemes. After evaluating the five schemes, we analyzed their differences and associations.

The rest of the paper is organized as follows: Sect. 2 introduces the emotion schemes. Section 3 proposes the evaluation framework. Section 4 conducts an annotation experiment and evaluates the five emotion schemes. Section 5 explores the differences and associations among these schemes. Section 6 is a discussion, and finally, Sect. 7 concludes the article.

Literature review

Emotion models

Emotion is a psychological phenomenon and has been studied from a variety of perspectives [9]. There are still many divergences in the understanding of emotion [10, 11]. Researchers have proposed a variety of emotion models. The discrete approaches propose that there are emotions that can be distinguished and that different types of emotions are independent. Dimensional approaches suggest that emotional states do not exist independently. There are multiple dimensions that make up the emotional space with smooth transitions between different emotions.

There are various discrete emotion models. Ekman identified that there are six basic emotions [12]. Scherer and Wallbott used seven major emotions in their cross-cultural questionnaire studies [13]. In ancient China, the theory of the "Seven Emotions" suggested that there are seven emotions [14, 15]. In Plutchik's wheel of emotions model [16], there are eight (four pairs) primary emotions, and each primary emotion is subdivided into three categories based on intensity. The combination of two neighboring primary emotions produces a complex emotion. There are also more refined emotion models, such as the OCC model, which adds 16 emotions to Ekman's six basic emotions for a total of 22 categories of emotions [17]. Parrott's three-layer structured emotion model has six primary emotions. Primary emotions are subdivided into secondary emotions, and secondary emotions are subdivided into tertiary emotions [18]. Cowen et al. found that there are 27 distinct varieties of emotional experience based on the self-reported method [19]. The classification of emotions by these models is shown in Table 1.

There are also various dimensional models. Russell's circumplex model has two dimensions: valence and arousal [20, 21]. Another important dimensional model is the PAD [22]. It has three dimensions: pleasure-displeasure, arousal-nonarousal, and dominance-submissiveness. Later, the PAD was extended to a valence-arousal-dominance (VAD) one [23]. There are also higher-dimensional models, such as the four-dimensional model [24] and the six-dimensional model [25].

Table 1 The emotion category in discrete models

Source	Number of emotions	Emotion taxonomy
Ekman [12]	6	happiness, sadness, anger, fear, disgust, and surprise
Scherer [13]	7	joy, fear, anger, sadness, disgust, shame, and guilt
Ancient China [14, 15]	7	joy, anger, sadness, fear, love, disgust, and desire
Plutchik [16]	8	joy-sadness, anger-fear, trust-disgust, surprise-anticipation.
Orthony (OCC) [17]	22	added 16 emotions to Ekman's six basic emotions: relief, envy, reproach, self-reproach, appreciation, shame, pity, disappointment, admiration, hope, fears-confirmed, grief, gratification, gloating, like, and dislike.
Parrott [18]	6	love, joy, anger, fear, sadness, and surprise
Cowen [19]	27	admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, and surprise.

Application of emotion models in text mining

The emotion lexicon and emotion-labeled dataset are the essential resources for textual emotion analysis. They use a wide variety of emotion models. The principles and process of emotion model selection are not described in detail in these resources. An emotion lexicon contains many emotion-related words, each assigned one or more emotion labels. There are many emotion lexicons. LIWC divides words into positive and negative and identifies three negative emotions: anxiety, anger, and sadness [26]. LIWC has been widely used in psychology and sociology but the categorization of emotions is not refined enough. The NRC lexicon uses Plutchik's eight emotions [27]. It provides fine-grained emotion with a limited ability to recognize polysemous words and implied emotions. WordNet-Affect adds hierarchically arranged emotion tags such as positive, negative, neutral, and ambiguous, and each tag is subdivided into a variety of emotions [28]. It performs fine-grained annotation of emotions and can capture the nuances of emotions. However, lexicon construction requires significant expertise.

An emotion-labeled dataset is a collection of data where each entry is labeled with the emotion category or affective state associated with it. There are many emotion-labeled datasets that use different emotion models. Bostan and Klinger surveyed those datasets [29] and found that most of them adopt discrete models, of which Ekman's and Plutchik's are the most frequent [4]. Some datasets use dimensional models, with VAD being the most popular. Others employ hybrid models [30], which label both emotion class and VAD scores.

Instead of using emotion models from the field of psychology, some datasets have customized emotion categories. For example, Grounded-Emotions uses only two emotions: happy and sad [31]. It has computational efficiency but affects the accuracy of emotion recognition. SemEval 2018 Task 1 uses four basic emotions: joy, sadness, fear, and anger [32]. EmoInt also follows the same four emotions [33]. The SemEval's four basic emotions have the advantages of simplicity, broad applicability, ease of assessment and comparison. However, this categorization method suffers from limited emotion categories. GoEmotions, which is a large emotion-labeled dataset, uses 27 emotions [34]. The use of 27 emotions has the advantage of fine-grained emotion categorization. However, it also increases annotation complexity, model complexity, and usage complexity.

Comparison of emotion classification schemes

If discrete emotion viewpoints are adopted for emotion analysis, it is necessary to choose an emotion classification scheme, which is needed for corpus annotation and model training and evaluation. Different emotion classification schemes vary in terms of psychological basis, number of emotions, set of emotions, etc. It has an impact on annotation as well as machine learning models in many aspects. Choosing an emotion classification scheme is not simple and deserves systematic study.

Some researchers have compared different models of emotion. Power et al. designed a questionnaire containing 30 emotion terms. A group of participants were asked how much in general they experienced each of the emotions. A confirmatory factor analysis was conducted to compare six different models of emotion with "goodness of fit" [35]. The purpose of the article was to analyze the quantities and relationships of basic emotions and not compare the commonly used emotion models.

A few studies have evaluated emotion classification schemes. Williams et al. compared six emotion classification schemes based on the ease of use of manual annotation and supervised machine learning performance [36]. The corpus was annotated separately using different emotion schemes, and then the six schemes were ranked using Inter-Annotator Agreement (IAA). Wood et al. conducted annotation experiments on tweets, comparing different annotation methods and emotion representation schemes [37]. They also use IAA as an evaluation indicator. Bruyne et al. conducted annotation experiments on tweets using the VAD model and compared different annotation methods (rating scales, pairwise comparison, and best-worst scaling) [38]. They evaluated the annotation methods based on the criterion of

inter-annotator agreement. They noticed the effects of different annotation methods on the time-consuming, complexity of affective judgments, but did not perform a comprehensive assessment of multiple metrics.

The use of an emotion classification scheme has important implications for the quality and efficiency of emotion analysis. How does one evaluate an emotion classification scheme? IAA is only an indicator of annotation reliability, and there are other aspects of annotation quality such as accuracy and coverage. For large corpus, annotation efficiency also needs to be considered. Therefore, a systematic assessment of emotion schemes from multiple perspectives is needed for emotion analysis.

This study developed a framework for evaluating emotion classification schemes. The framework can be applied to the evaluation of discrete emotion schemes. A good emotional scheme should lead to a balance between annotation quality and efficiency. Five criteria are proposed for this goal, which are solidity, coverage, agreement, compactness, and distinction. For the quantitative metrics, we designed a computational method based on annotation experiments. AHP was used to calculate the composite score. As an application of this framework, we collected Sina Weibo posts related to public events, evaluated five emotion classification schemes, ranked them, and analyzed the differences and associations of these schemes. This framework can evaluate emotion schemes from multiple aspects. It may be helpful to determine the emotion classification scheme in emotion analyses.

Methods

The evaluation framework

Choosing an emotion classification scheme suitable for the annotation of posts involves many qualitative and quantitative factors that need to be synthesized. AHP is an evaluation method that is capable of coping with both the rational and the intuitive to select the best candidates [39]. The elements of AHP include goals, criteria, metrics, and candidates. A goal is what is expected to be achieved. Criteria are refined based on the goal and then transformed into computable metrics. Candidates are the objects being evaluated.

What kind of emotion the text conveys is subjective, vague, and ambiguous. We conducted annotation experiments and obtained quantitative metrics from the annotation results. For the qualitative factors, pairwise comparisons were used to construct judgment matrices to quantify the qualitative issues. Finally, top-down weighting and addition were performed to obtain a composite score for each candidate.

Goal

The goal is to choose a suitable emotion classification scheme from a set of candidates, which should balance the quality and efficiency of annotation. A suitable scheme is expected to achieve a high quality of annotation. In addition, the efficiency of annotation also needs to be considered. Efficiency is the output that can be achieved with a given resource investment (time, manpower, budget, etc.). Text annotation is a resource-intensive and time-consuming task; efficient annotation can significantly reduce the time and cost.

Criteria

Emotion classification schemes may affect annotation quality and efficiency in many ways. Based on the goal of the evaluation, we identified five criteria as follows:

(1) Solidity

There are many emotion models, which may differ in their solidity. Solidity refers to their tightness and robustness in terms of logical structure, empirical validation, explanatory and predictive power, etc. For example, some models, such as Ekman's six emotions, have a great deal of scientific research, and some models have less scientific research. The use of a more solid model with accurate categorization of emotions, appropriate emotion granularity, and ease of understanding facilitates better annotation quality.

(2) Coverage

The categories in an emotion model should cover as many of the emotions in the corpus as possible. Coverage refers to how likely it is that a piece of text in the corpus contains a class of emotion that belongs to the emotion model. Insufficient coverage may cause some of the posts to lack appropriate emotion labels, resulting in mislabeling and lower annotation quality. Posts that are not labeled with emotion become ineffective outputs and reduce efficiency.

(3) Agreement

Posts are often labeled with multiple annotators. Different annotators may have different judgments about the emotions embedded in the post. Agreement refers to the degree of consistency between the annotation results of annotators. If multiple annotators select the same label for a post, the results are more reliable. Inter-annotator agreement is an important aspect of quality [40]. If the consistency is too low, the post will be difficult to use as ground truth, thus reducing the efficiency of the annotation.

(4) Compactness

Each scheme contains a number of emotion categories. Compactness refers to the number of emotions contained in an emotion scheme. Using a scheme with fewer emotion categories, the annotator uses less time and effort to make choices and is more efficient. If there are too many emotion categories, some may overlap, the annotator is prone to misuse. The burden of annotation work is greater.

(5) Distinction

It is crucial that each emotion category can be easily differentiated. Distinction refers to whether there is a clear distinction between the various emotions in the scheme. If there is a clear distinction, it is beneficial to reduce the cognitive load on the annotator and improve annotation efficiency.

Metrics

Assuming there are n posts in the corpus, each post is annotated by m annotators. The number of emotion schemes to be evaluated is v, each scheme contains $k_j,\ j = 1,2,\ldots,v$ categories of emotions.

(1) Metric of solidity

Solidity is subjective in nature. Obtaining quantized values of solidity is a complex issue. We used subjective judgments by consulting with people familiar with emotion models. Applying the AHP method to calculate the metric based on the pairwise comparisons. The elements of the judgment matrix are scaled from 1 to 9, and the computed solidity for each scheme ranges from 0 to 1.

(2) Metric of coverage

When using the emotion scheme j, each post was provided with k_j emotions and "neutral", "no suitable emotion", and "undistinguishable" options. If there are x_i posts labeled by annotator i with "no suitable emotion", the coverage is:

$$coverage = 1 - \frac{\sum_{i=1}^{m} x_i}{m \times n} \tag{1}$$

The coverage ranges from 0 to 1.

(3) Metric of agreement

IAA is an important measure of reliability [41]. The IAA is used as a measure to show how much the coders agree. Common metrics used include Scott's pi [42], Cohen's kappa [43], Fleiss' kappa [44], and Krippendorff's alpha [45]. Krippendorff's alpha is useful for multiple categories, multiple coders, can handle missing values, and corrects for randomness [46]. We employed Krippendorff's alpha (k-alpha) to assess agreement. We utilized Real Statistics software to compute both k-alpha and confidence intervals [47]. The k-alpha ranges from 0 to 1.

(4) Metric of compactness

The smaller the number of emotion categories contained in a scheme, the more compact it is. Among all the emotion schemes, the minimum number of categories is denoted as s_{min} and the maximum number of categories is denoted as s_{max}, the scheme j has k_j categories, its compactness is:

$$compactness = \frac{s_{max} - k_j + 1}{s_{max} - s_{min} + v} \tag{2}$$

The Laplace correction is used, with the numerator added 1 to avoid a compactness of 0. v is the number of emotion schemes, with the denominator adding v to avoid compactness > 1. The compactness ranges from 0 to 1.

(5) Metric of distinction

In the annotation process, the annotator is asked to give a unique emotion, and if he or she is not sure, he or she chooses the option "undistinguishable". If the annotator i judges that there are y_i posts "undistinguishable", then the distinction is:

$$distinction = 1 - \frac{\sum_{i=1}^{m} y_i}{m \times n} \tag{3}$$

The distinction ranges from 0 to 1.

Candidates

Researchers can choose any emotion classification scheme to evaluate for their own needs. In this paper, we chose five emotion schemes to demonstrate the application of the evaluation framework.

SemEval has four basic emotions: joy, sadness, anger, and fear. These four emotions are common to many basic emotion models.

Ekman's six basic emotions are happiness, sadness, anger, fear, disgust, and surprise. Ekman's six basic emotions have had a wide impact on psychology.

The Seven Emotions of China: joy, anger, sadness, fear, love, disgust, and desire. It has a long history and a wide influence in Eastern societies.

Plutchik's eight primary emotions are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation.

Pluchik's model of emotions has had a wide-ranging influence in psychology.

GoEmotions 27 emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, and surprise. GoEmotions is a recent large-scale emotion labeling dataset that attempts to cover the diverse types of emotions in web texts.

AHP model

Based on the analysis above, the AHP model is shown in Fig. 1.

There are six judgment matrices in the model. The importance of each criterion to the goal is subjective, and the goal-criterion matrix was obtained by pairwise comparisons expressed on a 1–9 scale. Similarly, quantitative comparison of candidates in terms of solidity is a complex problem. Pairwise comparison was also used to obtain the solidity-candidates matrix. The remaining four matrices were obtained by quantitative calculation. The coverage-candidates matrix was calculated based on the coverage of each scheme, with matrix element $a_{ij} = coverage_i/coverage_j$, i.e., the ratio of coverage of scheme i to scheme j. Similarly, the agreement-candidates matrix used the ratio of k-alpha, the compactness-candidates matrix used the ratio of compactness, and the distinction-candidates matrix used the ratio of distinction.

After the judgment matrix passed the consistency test, the priority vector of schemes was computed. The optimal scheme was selected based on their scores.

Data collection and cleaning

We used Octopus crawler to search for social event posts from Sina Weibo, with keywords including post-COVID-19 economy, health care reform, influenza A, negative population growth, college student employment, Russian-Ukrainian conflict, earthquakes, and GPT. The time range of microblog posting is 2022.12.1-2023.5.31, and a total of 15,098 microblogs were obtained.

A random sample of 3,000 posts was manually inspected to remove posts unrelated to social events. These include deleting posts with advertisement links, deleting posts with less than 5 words, and some posts that do not reflect the search intent, such as some posts under the keyword "earthquake" that have nothing to do with earthquakes, such as "pupil quake", which is just an Internet buzzword expressing surprise. After cleaning, we got 2,848 posts as a corpus. A few samples were taken from the corpus, as shown in Table 2.

We counted the number of words in each post, with a minimum of seven words, a maximum of 4743 words, and an average of 153 words. The distribution of word counts is shown in Fig. 2. Here, 58.4% of the posts had no more than 100 words, 88.5% had no more than 300 words, and 94.8% had no more than 500 words. Overall, while a small number of posts were long, most were short.

We also analyzed the number of sentences in each post. The mean of the number of sentences is 3.89, and the mode is 1. Here, 38.5% had only 1 sentence, 67.3% had no more than 3 sentences, 81.8% had no more than 5 sentences, and 91.2% had no more than 8 sentences. Most posts have a low number of sentences.

Manual annotation

We recruited five college students as annotators. They are all Chinese and have no religious beliefs. We provided

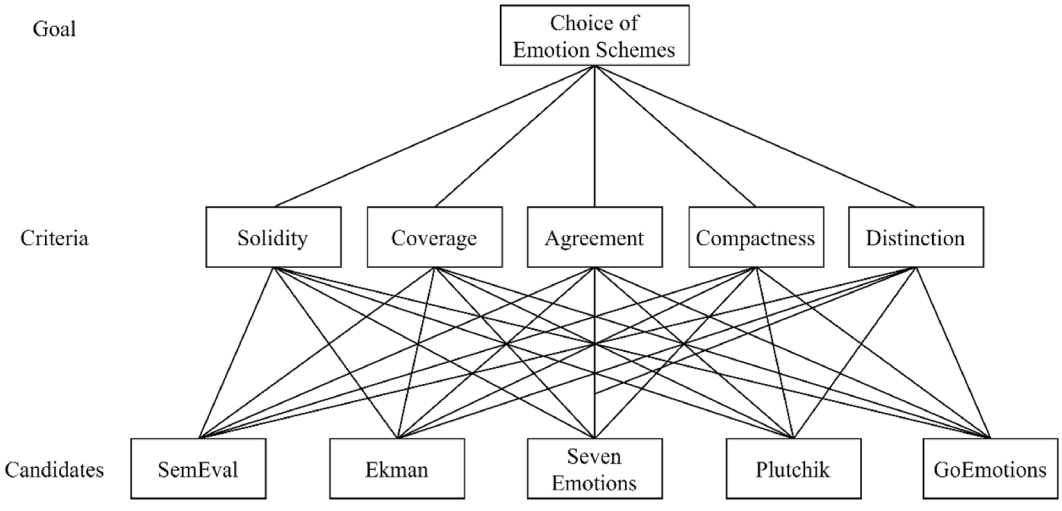

Fig. 1 The hierarchy model for the evaluation of the five emotion schemes

Table 2 Samples of Weibo posts

Keyword	Example Post
GPT	Recently, Chat GPT is being talked about everywhere, every teacher and professor you meet will talk about it, your classmates are talking about it when you walk on the road, and they are talking about it when you go back to your dormitory, it's a revolution in the industry!
Employment	Ten years ago, you could change your fate by graduating from secondary school, now a Ph.D. doesn't guarantee a satisfying job, and a real iron rice bowl is not a job that will support you for the rest of your life, but one that won't starve to death no matter where you go!
Earthquake	I slept like a dead pig. I didn't even know the sky was falling.
Economy	So satisfying to watch a play and eat with my mom. Adding to the post epidemic economy.
Influenza A	I don't understand how some people covering their ears and mouths means the virus doesn't exist. It doesn't matter if it's influenza A or the new coronary, it can kill people. If you're sick, take your medication, if it's serious, go to the hospital, so what's the point of coming to me and acting all shifty?
Population	Of the four provinces with negative resident population growth, Hebei decreased by 280,000, Hunan by 180,000, Henan by 110,000, and Shandong by 72,000. Among them, the natural population growth rates in Shandong and Henan were negative for the first time in recent decades.
Health care	As a family with two seniors over the age of 60, I fully support this health care reform!
International trade	The decoupling of China and the United States is accelerating, the retreat of foreign capital and the transfer of the industrial chain are very scary, and the situation of foreign trade is very serious!

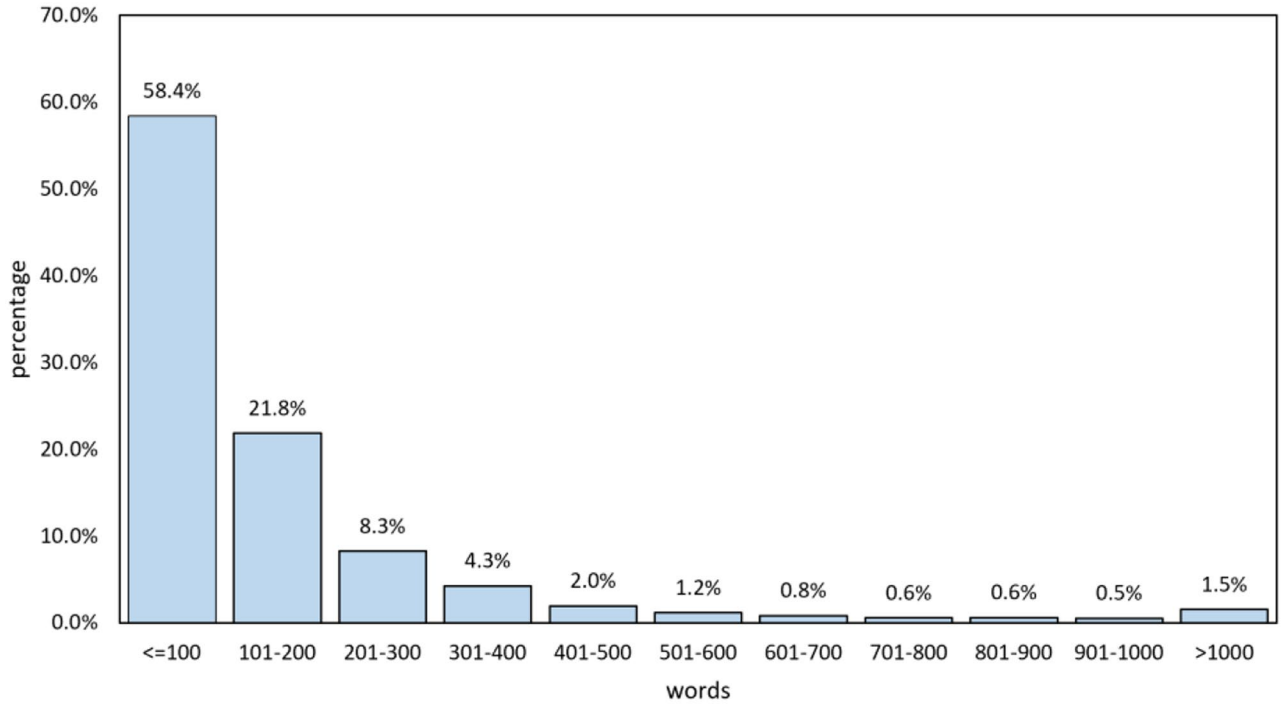

Fig. 2 Word count distribution of 2848 posts

an annotation guide, which includes an introduction to the annotation task, an introduction to the five emotion schemes, the meanings of various labels, and operation methods. The guide makes it clear that the emotions to be annotated are those of the post writers.

The annotators were first trained. The researchers explained the annotation guide to the annotators. Discussions were held with the annotator to solve their queries. Then 50 randomly selected posts from the corpus were pre-annotated, and the annotation results were discussed until a consensus understanding was achieved.

The labeling of each post includes the following 4 items:

(1) Select the emotion label. The options include all emotions in the current scheme, "neutral" means no emotion, "no suitable emotion" means there is no suitable emotion, and " undistinguishable" means that more than two emotions are difficult to distinguish.
(2) Degree of confidence, including 'sure' and 'not sure'.
(3) If 'not sure', explain why.

(4) Memo, describing what needs to be clarified in the labeling process.

Each post was annotated by the five annotators. The labeling process took one month and was divided into five phases. Five sequences were randomly generated for the five emotion schemes and assigned to annotators 1 ~ 5. For each phase, each annotator chose a particular emotion scheme to annotate according to the assigned sequence. At the end of each phase, the annotators had a three-day rest period to dilute the effect of the previous annotations on the subsequent annotations.

After the annotation was completed, the annotation results were collected and organized, the omissions and errors were manually checked, and the annotator was asked to correct them. Finally, the annotation results of the five emotion schemes were obtained.

Results

Annotation results

Distribution of emotions

The corpus contains 2,848 posts, and each post was labeled by five annotators. The labels include all the emotions in the current scheme, "neutral", "no suitable emotion", and "undistinguishable". All the labeling results of the 5 annotators were counted, and the percentage of each label was calculated. The percentages of each label under each emotion scheme are shown in Figs. 3, 4, 5, 6 and 7.

The distribution of emotions shows that the four basic emotions, sadness, joy (happiness), fear, and anger, have a relatively large share of the corpus, and their proportions are close to each other in each scheme, making them the main emotions in the corpus. However, GoEmotions disperses these four emotions into a variety of similar emotions due to the fine-grained division of emotions. In addition to the four main emotions, some emotions specific to each scheme, such as surprise and disgust in Ekman, disgust, love, and desire in Seven Emotions, and trust, anticipation, surprise, and disgust in Plutchik, although accounting for a relatively small proportion, are also present, indicating that the corpus covers all emotions.

Coverage and distinction

When the annotator cannot find an appropriate emotion in the current scheme, he or she selects "no suitable emotion", which is related to the metric of coverage. The other option, "undistinguishable", is used to compute the metric of distinction. Table 3 shows both options' percentages in each scheme.

The percentage of "no suitable emotion" varies significantly between the schemes, suggesting that the use of different schemes has a significant effect on coverage. While the percentage of "undistinguishable" is generally lower across all schemes, there are still some differences, suggesting that the differentiation may be slightly varied.

Agreement

Labeling a post by many annotators is like the voting process. The labels are candidates, and each label will have some votes, where at least one label has the largest number of votes, called the maximum number of votes (denoted max_votes), which can be used to simply

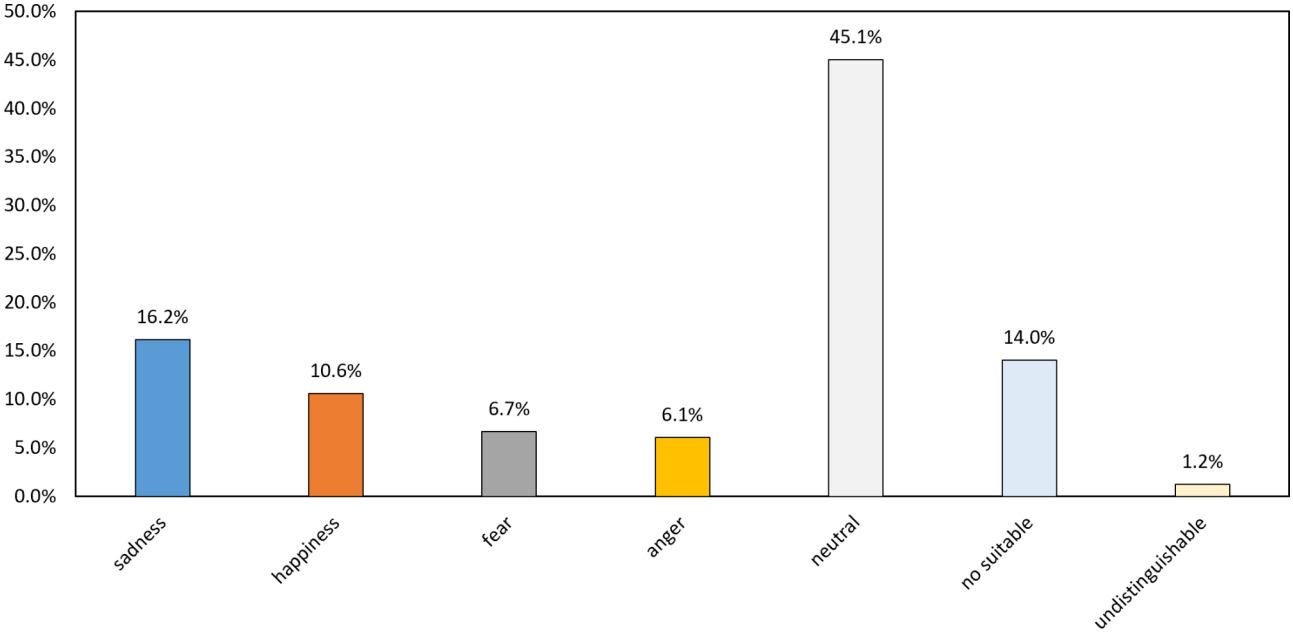

Fig. 3 Distribution of emotions based on the SemEval scheme

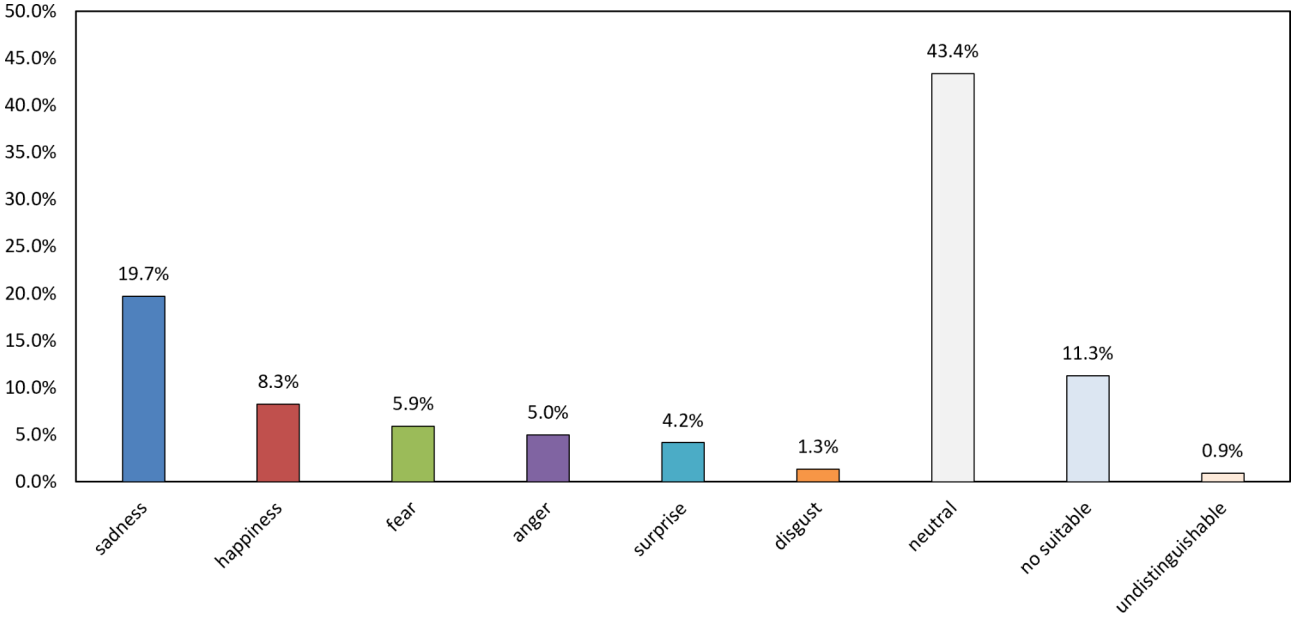

Fig. 4 Distribution of emotions based on the Ekman scheme

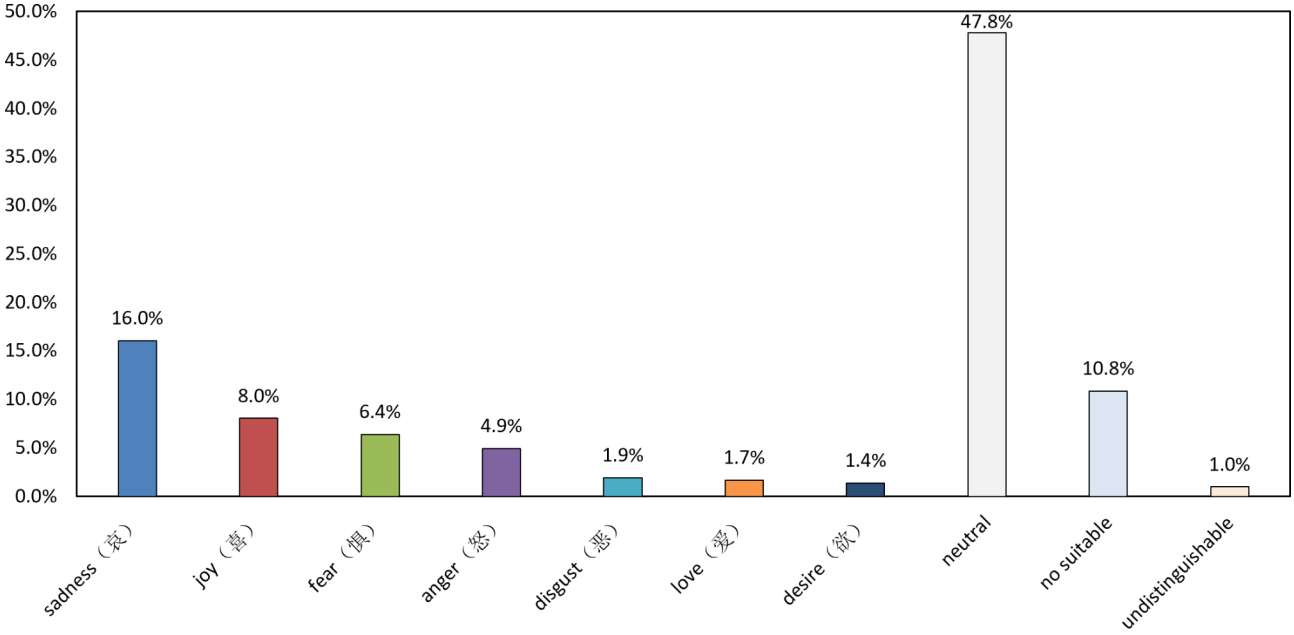

Fig. 5 Distribution of emotions based on the Seven Emotions scheme

reflect consistency. If every annotator chooses a different label, we denote it as NA (No Agreement), and the max_votes is 1. If $max\_votes \geq 3$, it can be considered a majority vote and achieve consensus. We calculated the percentages of posts with $max\_votes = 1$ and the ones with $max\_votes \geq 3$ for each scheme to indicate consistency. Krippendorff's alpha (k-alpha) was also used to measure the IAA. Table 4 shows the consistency of each scheme.

The data for NA ($max\_votes = 1$) show that there are a small number of posts that cannot be agreed upon in all schemes. The percentage of $max\_votes \geq 3$ is greater than 70% in all schemes, indicating that the annotators agree on the emotion embedded in most of the posts. The k-alpha of each scheme ranges from 0.33 to 0.41. In textual emotion labeling, k-alpha may be lower than the threshold of reliability [48]. Williams et al. reported that the range of k-alpha is 0.202 to 0.483 in their annotation work [36]. Despite the low k-alpha, there are some

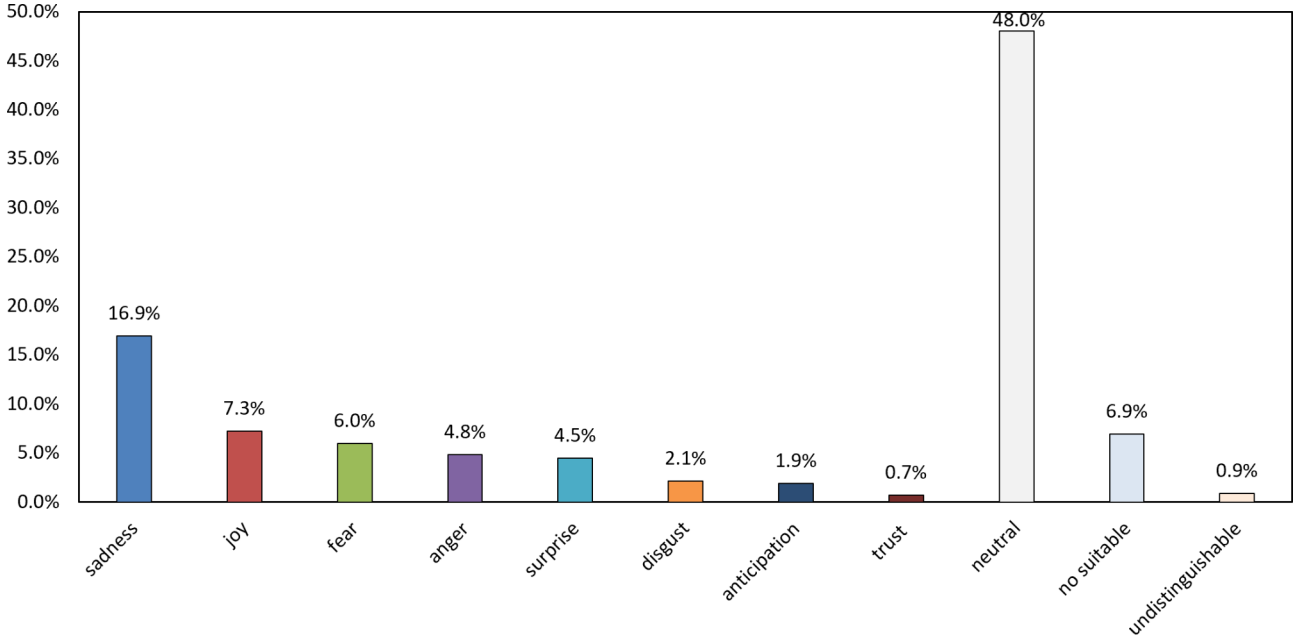

Fig. 6 Distribution of emotions based on the Plutchik scheme

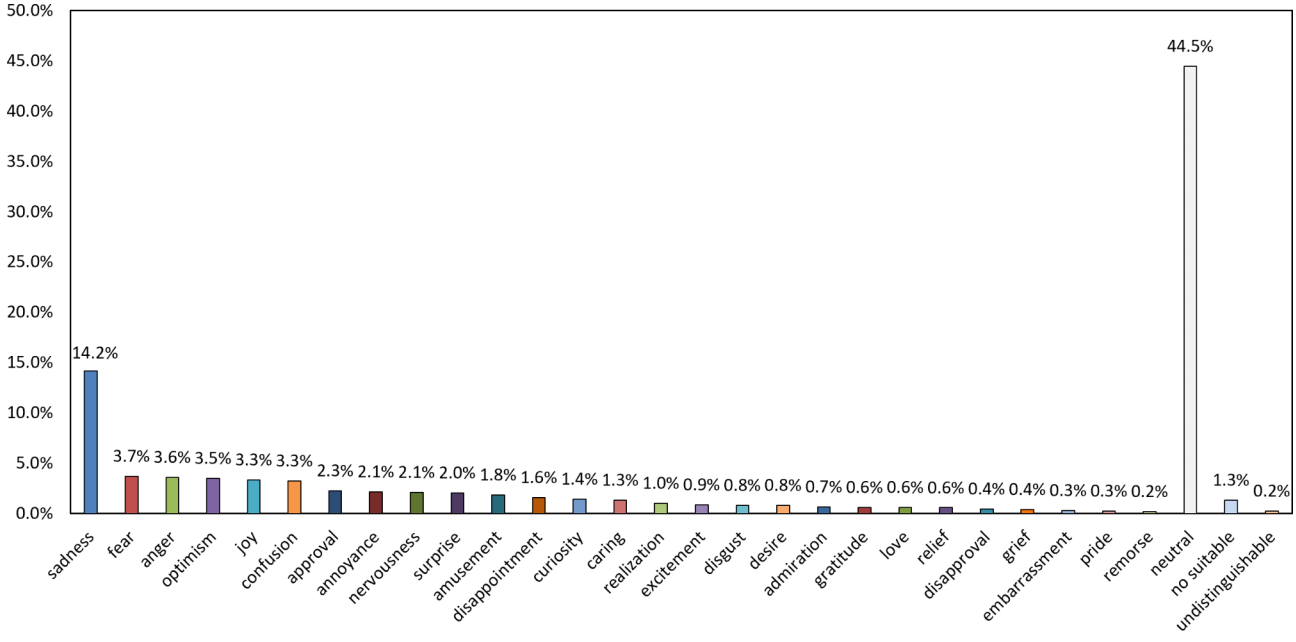

Fig. 7 Distribution of emotions based on the GoEmotions scheme

Table 3 Percentage of"no suitable emotion"and"undistinguishable"in each scheme

Emotion Schemes	no suitable emotion	undistinguishable
SemEval	14.03%	1.24%
Ekman	11.30%	0.91%
Seven Emotions	10.83%	0.98%
Plutchik	6.92%	0.86%
GoEmotions	1.34%	0.24%

Table 4 The measures of agreement in each scheme

Emotion Schemes	max_votes = 1	max_votes ≥ 3	k-alpha
SemEval	0.74%	79.92%	0.4067
Ekman	1.44%	73.63%	0.3332
Seven Emotions	1.83%	74.54%	0.3475
Plutchik	1.33%	75.49%	0.3613
GoEmotions	0.77%	78.97%	0.3952

Table 5 The goal-criteria matrix and the weights of each criterion

	Solidity	Coverage	Agreement	Compact	Distinction	Weight
Solidity	1	2	3	5	5	0.4020
Coverage	1/2	1	2	7	5	0.2962
Agreement	1/3	1/2	1	5	3	0.1780
Compact	1/5	1/7	1/5	1	2	0.0673
Distinction	1/5	1/5	1/3	1/2	1	0.0564

Table 6 The solidity-candidates matrix

	SemEval	Ekman	SevenEmotions	Plutchik	GoEmotions
SemEval	1	1/7	1	1/5	1/3
Ekman	7	1	5	3/2	5
SevenEmotions	1	1/5	1	1/3	1/2
Plutchik	5	2/3	3	1	3
GoEmotions	3	1/5	2	1/3	1

Table 7 The four metrics of each scheme

	Coverage	Agreement	Compactness	Distinction
SemEval	0.8597	0.4067	0.8571	0.9876
Ekman	0.8870	0.3332	0.7857	0.9909
SevenEmotions	0.8917	0.3475	0.7500	0.9902
Plutchik	0.9308	0.3613	0.7143	0.9914
GoEmotions	0.9866	0.3952	0.0357	0.9976

differences in k-alpha across the five schemes, implying that different schemes have an impact on consistency.

Evaluation results

We constructed all judgment matrices based on the AHP model. The importance of each criterion is subjective, and AHP supports the quantification of subjective judgments. By comparing the five criteria pairwise, the goal-criteria matrix was obtained, and weights were calculated, as shown in Table 5. The pairwise comparison was made on a scale of 1–9, with 1 indicating equal importance, 3 indicating moderate importance, 5 indicating strong importance, 7 indicating very strong importance, and 9 indicating extreme strong importance.

To obtain the exact value of the vector of weights $W = (w_1, w_2, \ldots, w_n)^T$, one needs to solve for $AW = \lambda_{max} W$, where A is the judgment matrix and λ_{max} is the largest eigenvalue. However, approximation methods are generally used. Here we used one of the common approximation algorithms: first normalizing the elements in each column of the judgment matrix, then averaging over each row, and finally normalizing.

Pairwise comparisons of the five schemes according to the solidity criterion yielded the solidity-candidates matrix, as shown in Table 6.

The other four matrices were obtained based on the metrics of coverage, agreement, compactness, and distinction, respectively. Coverage is equal to 1 minus the percentage of "no suitable emotion". Agreement is measured using the k-alpha. Compactness is obtained by substituting the number of categories in each scheme into Eq. (2). Distinction is equal to 1 minus the percentage of "undistinguishable". The four metrics of each scheme is shown in Table 7.

Table 8 The coverage-candidates matrix

Coverage-Candidates	SemEval	Ekman	Seven-Emotions	Plutchik	Go-Emotions
SemEval	1	0.9692	0.9641	0.9236	0.8714
Ekman	1.0318	1	0.9947	0.9529	0.8990
SevenEmotions	1.0372	1.0053	1	0.9580	0.9038
Plutchik	1.0827	1.0494	1.0438	1	0.9434
GoEmotions	1.1476	1.1123	1.1064	1.0600	1

Table 9 The agreement-candidates matrix

Agreement-Candidates	SemEval	Ekman	Seven-Emotions	Plutchik	Go-Emotions
SemEval	1	1.2206	1.1704	1.1257	1.0291
Ekman	0.8193	1	0.9588	0.9222	0.8431
SevenEmotions	0.8544	1.0429	1	0.9618	0.8793
Plutchik	0.8884	1.0843	1.0397	1	0.9142
GoEmotions	0.9717	1.1861	1.1373	1.0938	1

In the coverage-candidates matrix, the element a_{ij} is equal to the ratio of the coverage of scheme i to the coverage of scheme j. Similarly, the other three matrices are also calculated using the ratio of corresponding metrics. The four matrices are shown in Tables 8, 9, 10 and 11.

According to the five criteria-candidates' matrices, the performance of each scheme under each criterion can be calculated, as shown in Table 12.

To obtain the score of each scheme, we multiplied each weight of a scheme by the weight of its corresponding criterion, then added over all the criteria. Based on the scores, in descending order, the rankings are Ekman,

Table 10 The compactness-candidates matrix

Compactness-Candidates	SemEval	Ekman	SevenEmotions	Plutchik	GoEmotions
SemEval	1	1.0909	1.1429	1.1999	23.9998
Ekman	0.9167	1	1.0476	1.0999	22.0084
SevenEmotions	0.8750	0.9545	1	1.0499	21.0084
Plutchik	0.8333	0.9091	0.9524	1	20.0079
GoEmotions	0.0417	0.0455	0.0476	0.0500	1

Table 11 The distinction-candidates matrix

Distinction-Candidates	SemEval	Ekman	SevenEmotions	Plutchik	GoEmotions
SemEval	1	0.9967	0.9974	0.9962	0.9900
Ekman	1.0033	1	1.0007	0.9995	0.9933
SevenEmotions	1.0026	0.9993	1	0.9988	0.9926
Plutchik	1.0038	1.0005	1.0012	1	0.9938
GoEmotions	1.0101	1.0068	1.0075	1.0063	1

Plutchik, GoEmotions, Seven Emotions, and SemEval, as shown in Fig. 8.

Each scheme has its own strengths and weaknesses, and they differ significantly in terms of solidity, coverage, agreement, and compactness. The performance of each scheme under each criterion was visually compared in Fig. 9.

As shown in Fig. 9, Ekman and Plutchik are better at solidity, GoEmotions and Plutchik are better at coverage, SemEval and GoEmotions are better at agreement, and SemEval and Ekman are better at compactness. There is very little difference in distinction, as the percentage of "undistinguishable" posts is close across each scheme.

Sensitivity analysis

Criteria weights were determined based on subjective judgment, with ambiguity and randomness. Do changes in the criteria weights have a significant impact on the scoring results? We performed a sensitivity analysis of the weights of the criteria. The sum of the weights of each criterion was set to 1, and the weights of only one

Table 12 The performance of the five schemes under each criterion

	Solidity	Coverage	Agreement	Compactness	Distinction
SemEval	0.0600	0.1887	0.2206	0.2727	0.1992
Ekman	0.4470	0.1947	0.1807	0.2500	0.1999
SevenEmotions	0.0765	0.1957	0.1885	0.2386	0.1997
Plutchik	0.2896	0.2043	0.1959	0.2273	0.2000
GoEmotions	0.1269	0.2166	0.2143	0.0114	0.2012

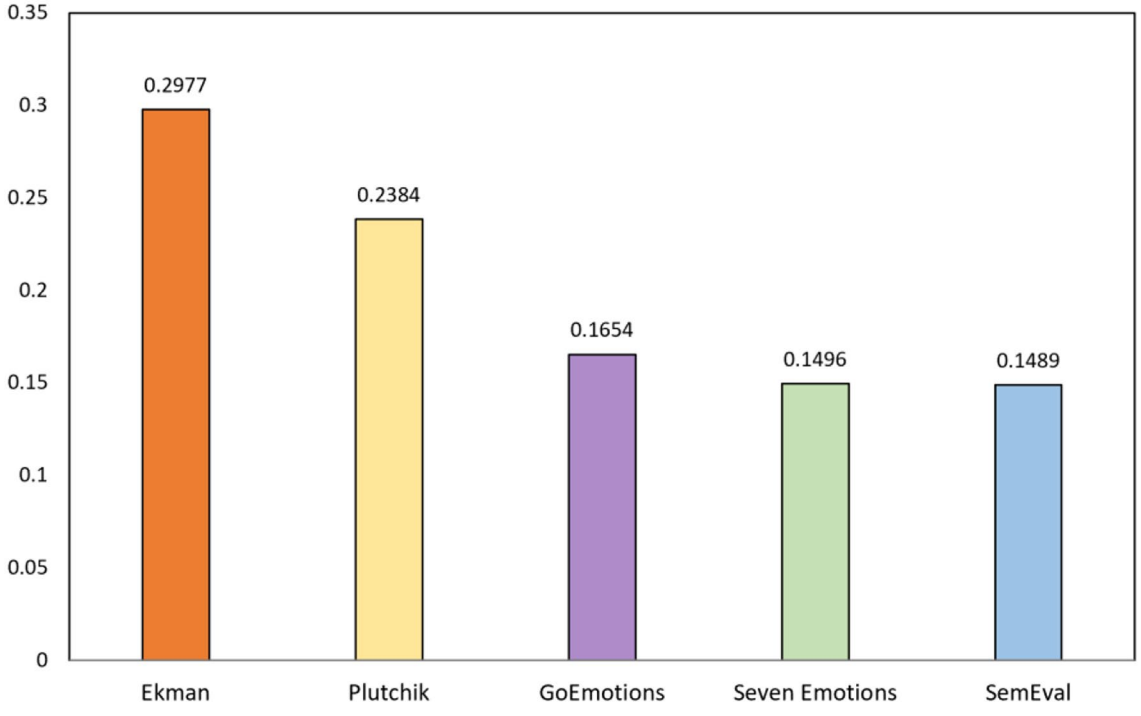

Fig. 8 The scores of the five emotion schemes

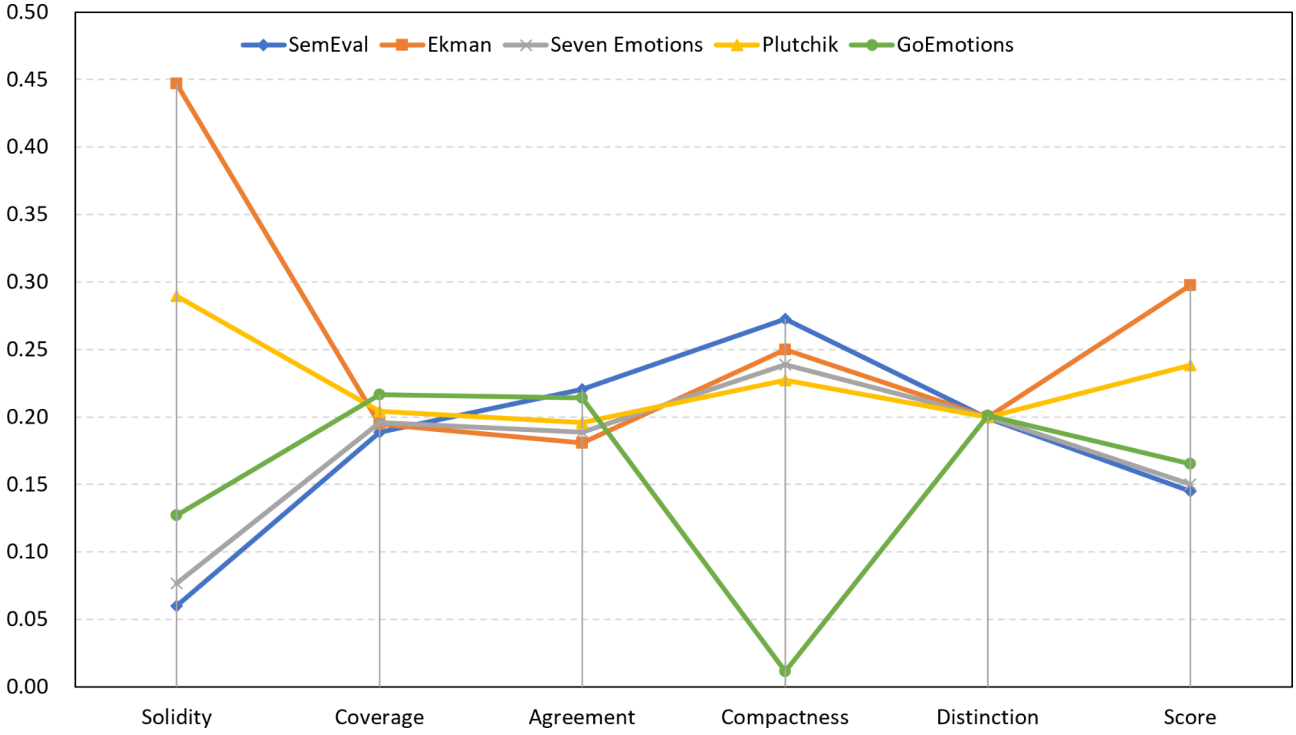

Fig. 9 Performance of the five emotion schemes under the five criteria

criterion were adjusted individually at a time, while the weights of the remaining criteria were distributed in equal proportions to the initial weights.

The initial weight of solidity is 0.4020, and when it varies within [0.0375, 1], Ekman and Plutchik remain in the 1st and 2nd places, and the ranking results are basically unchanged. When the coverage varies within [0, 0.9016], the ranking result is basically unchanged. When agreement varies within [0, 0.8313], the result is basically unchanged. Compactness varies within [0, 0.8750], and the results are almost the same. Ranking results are not sensitive to changes in distinction.

When the weights of each criterion change in a wide range, the ranking results remain essentially unchanged, indicating that the results are robust. Of course, only the change of single criterion weights was discussed here, and the case of multiple weights changing at the same time was not discussed.

Comparison of different schemes

According to the evaluation results, the Ekman scheme scored the highest. It is better in terms of solidity and compactness but not in terms of coverage and agreement. Whether it is ideal and how similar or different it is from others needs to be analyzed in depth.

Similarities

Each of the five schemes has pros and cons; which of them is more similar? Firstly, the results of the five annotators are aggregated by majority vote. For a post, if $max\_votes < 3$, the consensus cannot be achieved by a majority vote, it is denoted as NC (No Consensus). For two emotion schemes, the NC co-occurrence of all posts was used to measure the similarity between them. The NC co-occurrence matrix is shown in Table 13, where the diagonal elements a_{ii} are the proportion of NC posts under scheme i, and the other elements a_{ij}, $i \neq j$ are the proportion of posts that are NC in both scheme i and scheme j.

Table 13 No Consensus Co-occurrence Matrix

	SemEval	Ekman	Seven Emotions	Plutchik	GoEmotions
SemEval	19.8%	12.6%	12.6%	11.9%	8.7%
Ekman	12.6%	26.4%	17.2%	18.5%	10.7%
Seven Emotions	12.6%	17.2%	25.5%	17.9%	11.0%
Plutchik	11.9%	18.5%	17.9%	24.5%	10.7%
GoEmotions	8.7%	10.7%	24.5%	10.7%	21.0%

The similarity of the five schemes cannot be directly observed from the NC co-occurrence matrix. Multidimensional scaling (MDS) is a dimensionality reduction and visualization method that could map high-dimensional data to lower dimensions, while keeping the distance relationship between data points, and facilitating observation of patterns in the data. We employed PROXSCAL for MDS to demonstrate the similarity

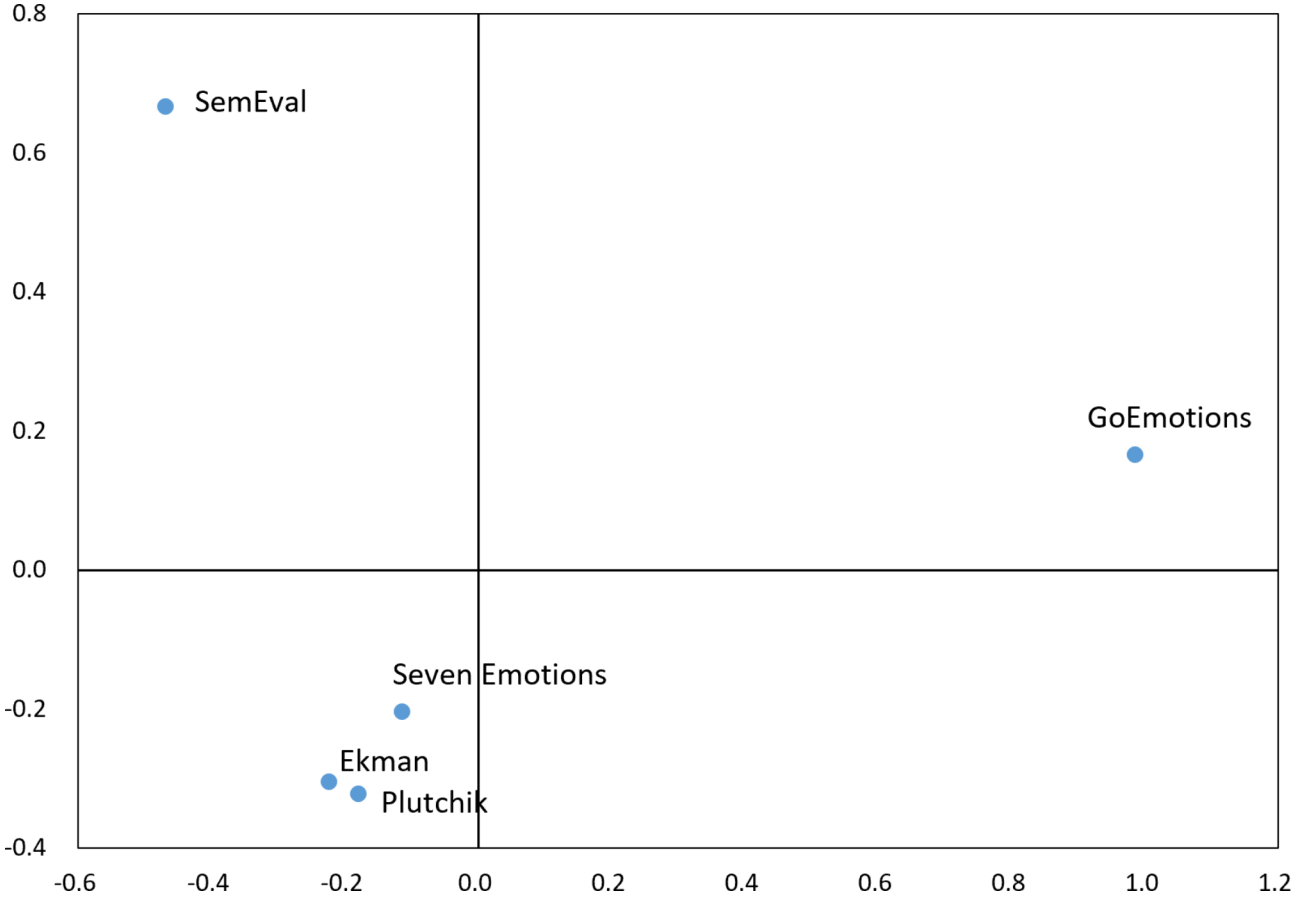

Fig. 10 Multidimensional scaling of the five emotion schemes

Table 14 Contingency table of SemEval and Ekman

		Ekman						
		sadness	**surprise**	**anger**	**happiness**	**fear**	**disgust**	**no suitable**
SemEval	sadness	455	6	6	6	17	0	13
	anger	16	4	126	2	0	6	3
	joy	16	7	3	236	0	2	30
	fear	24	4	0	0	100	0	8
	no suitable	31	72	9	9	16	3	44

of the schemes in 2D space, as shown in Fig. 10. Here, stress=0.0305 and D.A.F=0.99844, lower stress (to a minimum of 0) and higher D.A.F (to a maximum of 1) indicate that the fit is good in two dimensions. The depiction is highly explanatory. In 2D space, the more similar the schemes, the closer they are to each other.

In Fig. 10, GoEmotions is separated from the other schemes by dimension 1, while dimension 2 separates SemEval and GoEmotions from the remaining three schemes. According to spatial proximity, the five schemes can be grouped into three: Ekman, Plutchik, and the Seven Emotions form a group; SemEval is one group; and GoEmotions is also a group. This suggests that Ekman, Plutchik, and Seven Emotions are more alike. The similarities here are only those analyzed from the perspective of NC co-occurrence.

Correspondence analysis

While the Ekman scheme scored the highest, it was not the most dominant in terms of coverage and agreement. Correspondence analysis was conducted between Ekman's and others to examine the association between emotion categories in different schemes.

(1) Ekman and SemEval.

These posts with $max\_votes \geq 3$ could get the final label by majority vote. The number of posts in the two

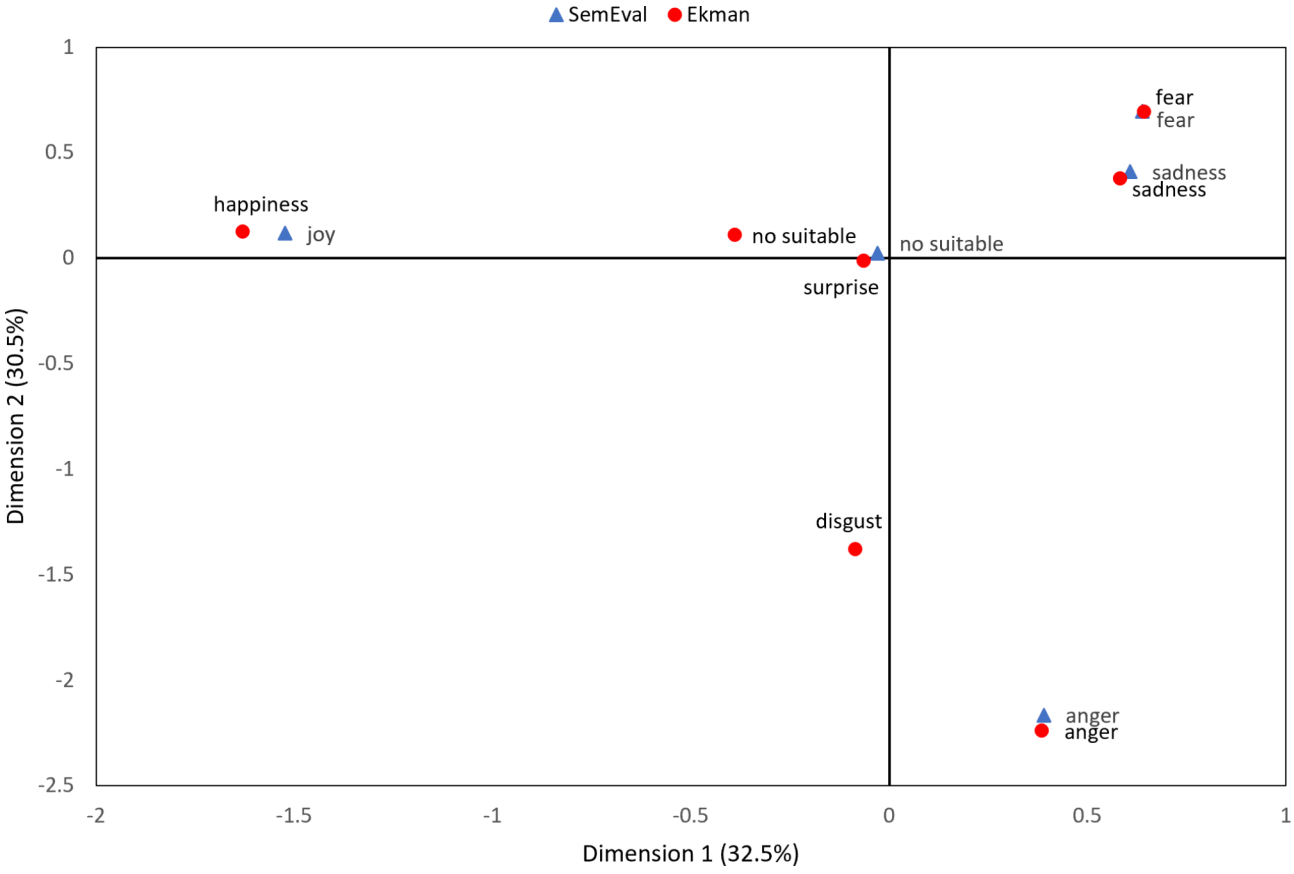

Fig. 11 Ekman (red dot) versus SemEval (blue triangle)

schemes was counted to get the contingency table, as shown in Table 14.

The correspondence analysis result is shown in Fig. 11, where the explained variance of the two dimensions is 62.5%. Fear, sadness, and anger are consistently linked, while joy correlates closely with happiness. The emotions of surprise and disgust in Ekman do not have a strong association with any emotion in SemEval, indicating their necessity.

(2)Ekman and the Seven Emotions.

The correspondence analysis result is presented in Fig. 12. The explained variance of the two dimensions is 60.2%. Dimension 1 separates positive emotions (joy, love, and desire) from negative emotions (anger, sadness, fear, and disgust).

The four synonymous emotions (fear, sadness, disgust, and anger) are highly consistent. In Seven Emotions, joy is strongly linked to happiness in Ekman's model, while love is strongly linked to happiness, and desire has no obvious counterpart in Ekman. On the other hand, Seven Emotions does not have a clear corresponding emotion for surprise as in Ekman.

(3)Ekman and Plutchik.

The correspondence analysis result is shown in Fig. 13. The five synonymous emotions (sadness, disgust, anger, fear, and surprise) are highly consistent. Plutchik's joy is highly consistent with Ekman's happiness. Compared to Ekman, Plutchik has two more positive emotions: trust and anticipation, trust is somewhat associated with happiness, and anticipation is weakly associated with happiness.

(4)Ekman and GoEmotions.

The correspondence analysis result is shown in Fig. 14. Each emotion of Ekman corresponds to a number of similar emotions in GoEmotions, such as: happiness corresponds to joy, admiration, excitement, and gratitude; anger corresponds to anger; and annoyance; sadness corresponds to sadness, grief, and remorse; fear corresponds to fear and nervousness; disgust corresponds to disgust; and surprise corresponds to surprise and curiosity.

However, some emotions in GoEmotions are not clearly related to Ekman, such as approval, disapproval, love, pride, amusement, embarrassment, desire, and caring. This indicates the independence of these emotions.

The above analyses show that the correspondence between emotions in each scheme is complex. Ekman is more similar to Seven Emotions and Plutchik, and their synonymous emotions, sadness, fear, anger, and disgust, are highly consistent. However, the only positive emotion in Ekman is happiness, which corresponds to joy in

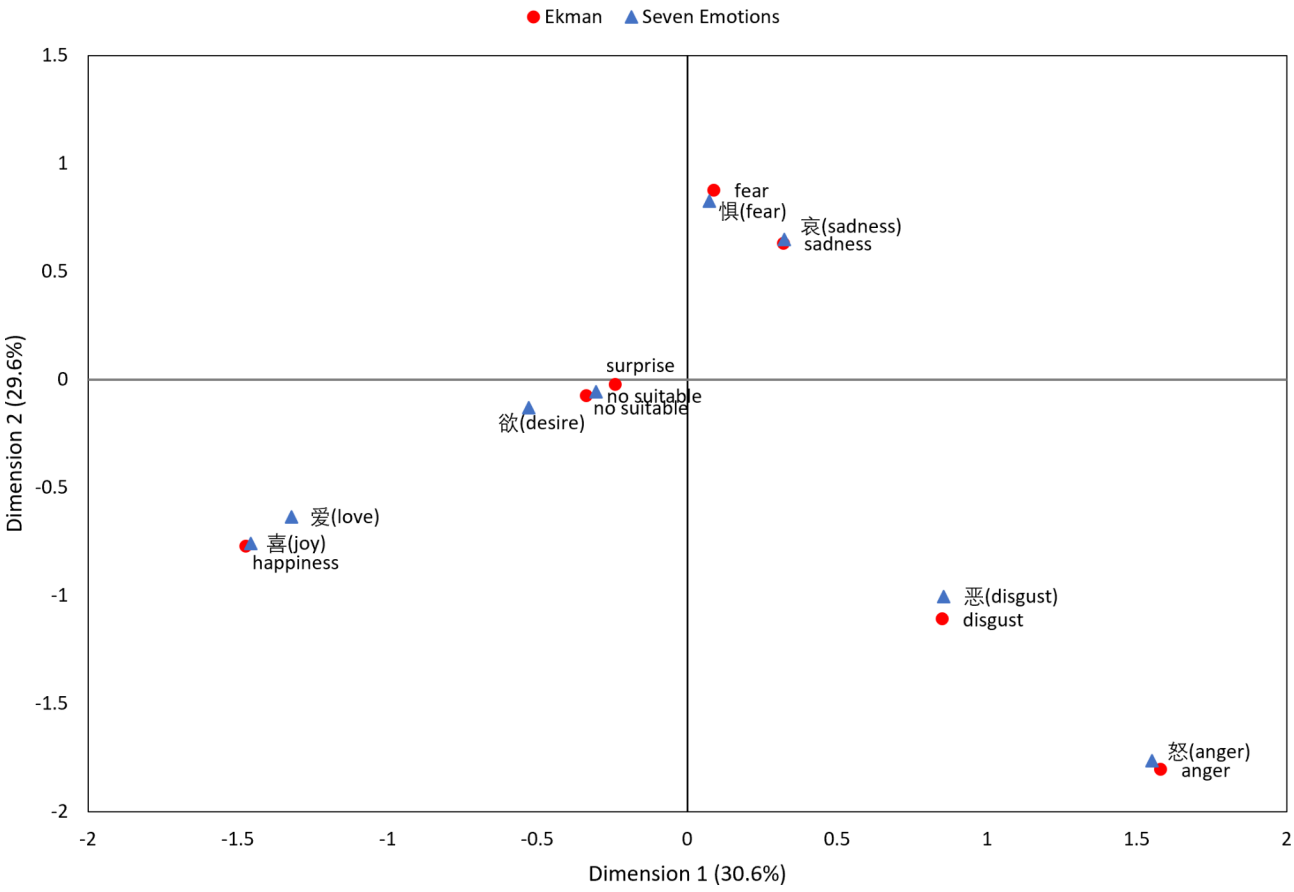

Fig. 12 Ekman (red dot) versus Seven Emotions (blue triangle)

Seven Emotions and Plutchik, and there are more positive emotions that are not expressed in Ekman. These positive emotions in Seven Emotions and Plutchik have a non-negligible proportion in the corpus. For the current corpus, there is a lack of sufficient positive emotion categories in the Ekman scheme.

Discussions

How to choose an appropriate one from the many emotion schemes is an important problem in emotion analysis. This paper makes two contributions to the selection of discrete emotion schemes. First, an evaluation framework was proposed to provide an integrated assessment of multiple factors, which helps to select an overall better scheme and overcome the shortcomings of a single indicator. Second, five commonly used emotion schemes were evaluated to select the scheme with the highest score. These five schemes were analyzed for similarities and differences, which helped to provide insight into the strengths and weaknesses of each scheme.

The evaluation framework consists of five criteria, including solidity, coverage, agreement, compactness, and distinction. Solidity reflects the credibility of the emotion scheme. Coverage reflects the completeness of emotional categories. Agreement reflects the consistency of the annotators. Compactness reflects the degree of non-redundancy in emotion categories. Distinction reflects whether there are significant differences between emotions. These criteria are key factors that affect the quality and efficiency of emotion analysis. Previous studies [36, 37] have used a single indicator, IAA, which is the agreement criterion in the framework. The IAA is an indicator of annotation reliability. It reflects only one aspect of annotation quality and cannot cover other quality indicators, nor does it reflect annotation efficiency. This is supported by the results of our evaluation. If only the IAA was used for ranking, the priority would be SemEval, GoEmotions, Plutchik, Seven Emotions, and Ekman. However, this rank is not consistent with the other four criteria, and the final scores differ. This implies that agreement cannot cover the other four criteria.

We evaluated the five commonly used emotion schemes and found that Ekman scored the highest. This is close to the actual adoption in the field of textual emotion analysis [4]. According to our evaluation, the Ekman scheme has stronger evidence and the highest solidity score. It also performed better in terms of compactness. However, the Ekman scheme has limited coverage, which

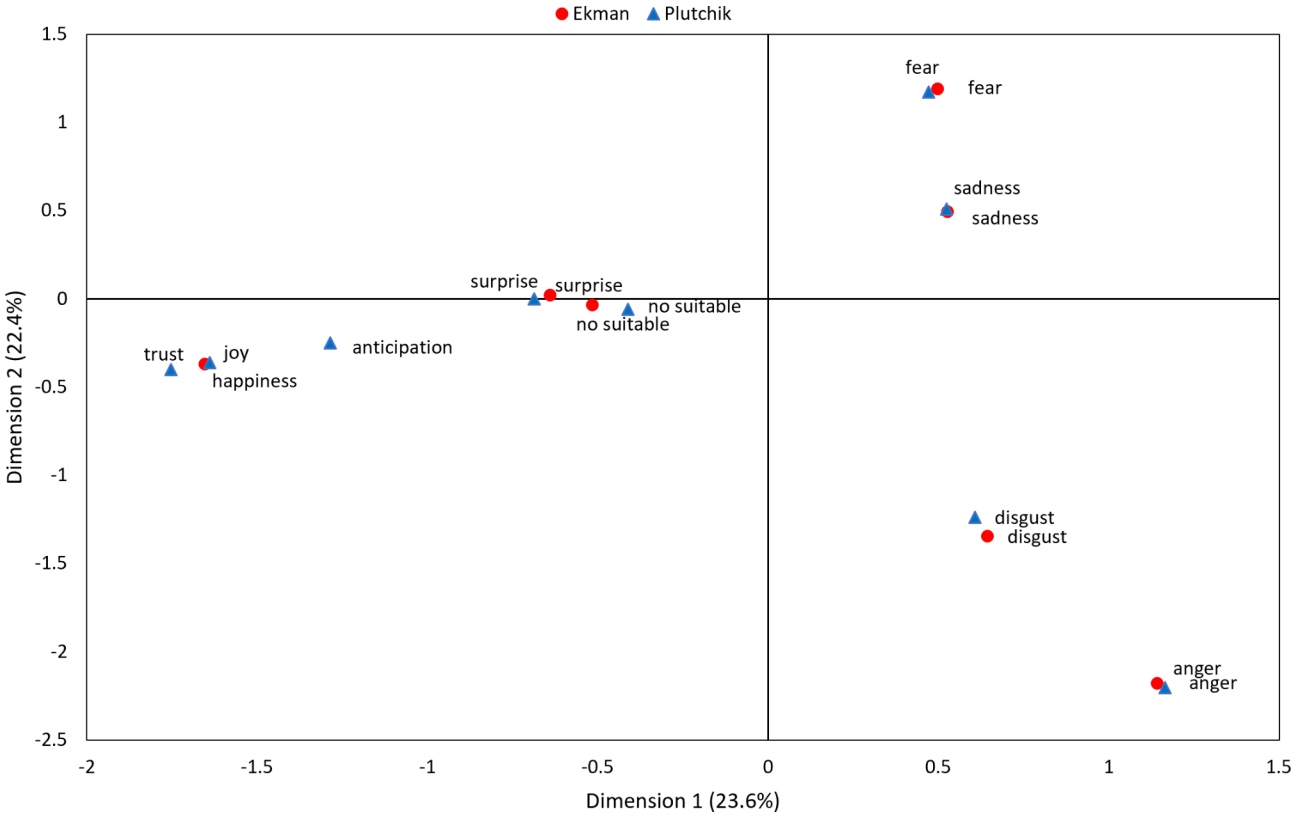

Fig. 13 Ekman (red dot) versus Plutichik (blue triangle)

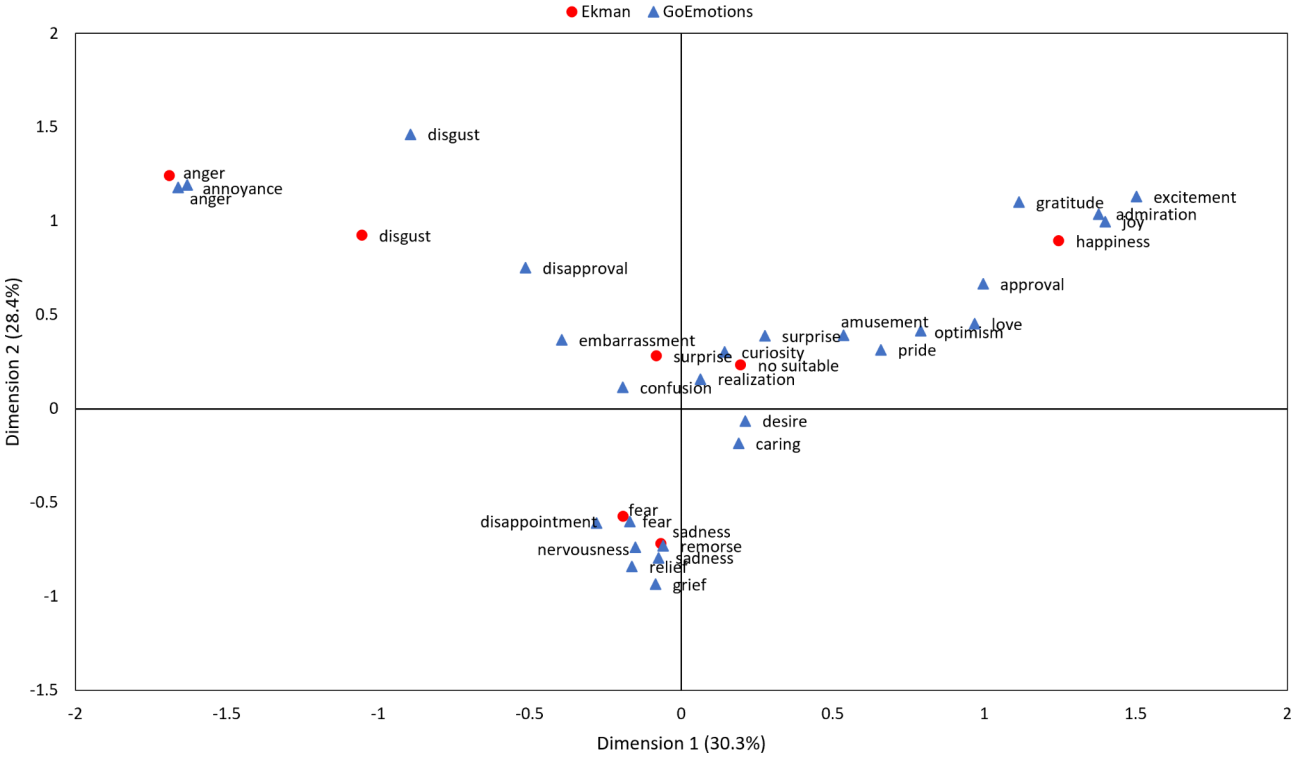

Fig. 14 Ekman (red dot) versus GoEmotions (blue triangle)

is supported by findings from Williams [36]. The reason may be the absence of positive emotion categories, and some studies have argued that there may be more than six basic emotions [49]. The addition of emotion categories does not always decrease agreement. For example, GoEmotions includes 27 emotions and still has a high level of agreement.

This study has limitations. Many factors may impact emotion annotation, including emotion classification schemes, annotators, the annotation process, emotion ontology, single or multi-labels, etc. This study only considered the effects of the emotion classification scheme. We selected only five emotion schemes for the evaluation, not all of them, and the ranking results are limited. In addition, textual emotion is domain specific. Emotional distribution may vary in different corpora. This study used Weibo posts focused on public events. Using a different corpus, the evaluation results may vary.

Conclusions and future works

Emotion analysis of text requires the selection of an emotion model. There are many discrete emotion schemes that need to be evaluated from multiple perspectives. This paper proposed an evaluation framework with the goal of achieving a balance between quality and efficiency in emotion analysis, for which five criteria are identified, which are solidity, coverage, agreement, compactness, and distinction. Indicators were designed for each criterion, with quantitative indicators calculated from the results of annotated experiments and qualitative indicators using pairwise comparisons. The AHP method was used to realize the combination of qualitative and quantitative metrics. As an application of this framework, Weibo posts in the domain of public events were collected, and five emotion classification schemes were evaluated. The results of the evaluation show that the Ekman scheme is the best, but it is deficient in coverage and agreement. The Ekman scheme has only one positive emotion, happiness, which may lead to less accurate labeling results for positive emotion texts.

In recent years, deep learning has developed rapidly and has advantages in emotion analysis. Commonly used deep learning models include CNN, LSTM, Bi-LSTM, GRU, and transformer-based models. CNNs can efficiently capture local features and are suitable for emotion analysis of short texts. LSTMs retain sequential information through a gating mechanism, and Bi-LSTM goes a step further by processing sequential and reverse-ordered contexts through bi-directional propagation, which enhances the comprehensive understanding of emotion. GRU serves as a variant of LSTM that reduces computational complexity while retaining similar performance. Transformer-based models such as BERT, XLM, and GPT capture long-distance dependencies through a self-attentive mechanism. BERT performs well in multiple emotion categorization tasks, while XLM demonstrates its cross-linguistic power in multilingual emotion analysis, and GPT performs well in emotion generation and comprehension tasks.

Deep learning models perform well in emotion analysis, but manual selection of an emotion classification scheme is still crucial. This is because emotion is a complex psychological phenomenon that is not fully understood. The appropriate emotion model needs to be selected based on the purpose of application. Different scenarios have specific needs for emotion classification and manually determining the emotion scheme allows for customized adjustments. Emotions in text are implicit and models need to be trained with annotated data. The emotion classification scheme needs to be determined while annotating the data. Therefore, the emotion scheme evaluation method still has value.

In future research, the proposed framework will be expanded to integrate both discrete and continuous emotion schemes. This expansion will likely require modifications to the existing metrics and their corresponding computational methodologies. Furthermore, we will investigate the quantification of some criteria within the framework. For instance, we will explore the use of bibliometric techniques to measure the solidity of emotion schemes. Finally, we aim to extend the framework's applicability by evaluating a broader range of emotion schemes across various domains. This will enable a comprehensive analysis of how domain-specific characteristics influence the selection of emotion schemes.

Author contributions

FZ contributed to the conception, design, analysis of the manuscript. JC contributed to data collection and analysis. QT contributed the results and discussions. YT contributed to the evaluation of emotion schemes. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the National Natural Science Foundation of China (Grant No. 71571190), the Guangdong Province Key Research Base of Humanities and Social Sciences (Grant No. 2022WZJD012), and Key Issues on High-Quality Development of the Guangdong-Hong Kong-Macao Greater Bay Area (Grant No. XK-2023-007).

Data availability

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 27 May 2024 / Accepted: 17 September 2024
Published online: 27 September 2024

References

1. Mäntylä MV, Graziotin D, Kuutila M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Comput Sci Rev. 2018;27:16–32.
2. Rodríguez-Ibánez M, Casánez-Ventura A, Castejón-Mateos F, Cuenca-Jiménez. P-M. A review on sentiment analysis from social media platforms. Expert Syst Appl. 2023;223:119862.
3. Deng J, Ren F. A survey of textual emotion recognition and its challenges. IEEE Trans Affect Comput. 2023;14:49–67.
4. Acheampong FA, Wenyu C, Nunoo-Mensah H. Text-based emotion detection: advances, challenges, and opportunities. Eng Rep. 2020;2.
5. Zhang F, Tang Q, Chen J, Han N. China public emotion analysis under normalization of COVID-19 epidemic: using Sina Weibo. Front Psychol. 2023;13:1066628.
6. Xu C, Zheng X, Yang F. Examining the effects of negative emotions on review helpfulness: the moderating role of product price. Comput Hum Behav. 2023;139:107501.
7. Harmon-Jones E, Harmon-Jones C, Summerell E. On the importance of both dimensional and discrete models of emotion. Behav Sci. 2017;7:66.
8. Beck J. Quality aspects of annotated data. AStA Wirtsch Sozialstat Arch. 2023;17:331–53.
9. Lerner JS, Li Y, Valdesolo P, Kassam KS. Emotion and decision making. Annu Rev Psychol. 2015;66:799–823.
10. Kuppens P. Improving theory, measurement, and reality to advance the future of emotion research. Cognition Emot. 2019;33:20–3.
11. Brady M, Précis. Emotions: the basics. J Philos Emot. 2021;3:1–4.
12. Ekman P. An argument for basic emotions. Cognition Emot. 1992;6:169–200.
13. Scherer KR, Wallbott HG. Evidence for universality and cultural variation of differential emotion response patterning. J Pers Soc Psychol. 1994;66:310–28.
14. Confucius. The Book of rites (Li Ji). Createspace Independent Pub; 2013.
15. Wang Y. The three-character classic. People's Literature Publishing House; 2020.
16. Plutchik R. The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci. 2001;89:344.
17. Ortony A, Clore GL, Collins A. The cognitive structure of emotions. Cambridge, MA: Cambridge University Press; 1990.
18. Parrott WG. Emotions in social psychology: key readings. Psychology; 2001.
19. Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc Natl Acad Sci. 2017;114:E7900–9.
20. Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39:1161–78.
21. Posner J, Russell JA, Peterson BS. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol. 2005;17:715–34.
22. Russell JA, Mehrabian A. Evidence for a three-factor theory of emotions. J Res Pers. 1977;11:273–94.
23. Bradley MM, Lang PJ. Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry. 1994;25:49–59.
24. Izard CE. The psychology of emotions. Springer Science & Business Media; 1991.
25. Frijda NH. The emotions. Cambridge University Press; 1986.
26. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol. 2010;29:24–54.
27. Mohammad SM, Turney PD. Crowdsourcing a word–emotion association lexicon. Comput Intell. 2013;29:436–65.
28. Strapparava C, Valitutti A. WordNet-Affect: an Affective Extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon; 2004. pp. 1083–6.
29. Bostan L-A-M, Klinger R. An analysis of annotated corpora for emotion classification in text. In: Proceedings of the 27th International Conference on Computational Linguistics. 2018. pp. 2104–2119.
30. Park EH, Storey VC. Emotion ontology studies: a framework for expressing feelings digitally and its application to sentiment analysis. ACM Comput Surv. 2023;55:1–38.
31. Liu V, Banea C, Mihalcea R, Grounded. emotions. 2017 Seventh Int Conf Affect Comput Intell Interact (ACII). 2017;:477–83.
32. Mohammad S, Bravo-Marquez F, Salameh M, Kiritchenko S. SemEval-2018 task 1: affect in tweets. Proc 12th Int Work Semantic Evaluation. 2018;:1–17.
33. Mohammad S, Bravo-Marquez F. WASSA-2017 shared task on emotion intensity. In: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2017. pp. 34–49.
34. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. Proc 58th Annu Meet Assoc Comput Linguistics. 2020;:4040–54.
35. Power MJ. The structure of emotion: an empirical comparison of six models. Cogn Emot. 2006;20:694–713.
36. Williams L, Arribas-Ayllon M, Artemiou A, Spasic I. Comparing the utility of different classification schemes for emotive language analysis. J Classif. 2019;36:619–48.
37. Wood ID, McCrae JP, Andryushechkin V, Buitelaar P. A comparison of emotion annotation approaches for text. Information. 2018;9:117.
38. Bruyne LD, Clercq OD, Hoste V. Annotating affective dimensions in user-generated content. Lang Resour Eval. 2021;55:1017–45.
39. Saaty TL, Vargas LG, Models, Methods C. Applications of the Analytic Hierarchy process. Int Ser Oper Res Manag Sci. 2012. https://doi.org/10.1007/978-1-4614-3597-6.
40. Braylan A, Alonso O, Lease M. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In: Proceedings of the ACM Web Conference 2022. 2022. pp. 1720–30.
41. Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. Hum Commun Res. 2004;30:411–33.
42. Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opin Q. 1955;19:321.
43. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
44. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378–82.
45. Krippendorff K. Bivariate agreement coefficients for reliability of data. Sociol Methodol. 1970;2:139.
46. Krippendorff K. Content analysis: An Introduction to its methodology. 4th Edition. Sage publications; 2019.
47. Zaiontz. Real Statistics using Excel. 2020. www.real-statistics.com
48. Antoine J-Y, Villaneau J, Lefeuvre A. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. Proc 14th Conf Eur Chapter Assoc Comput Linguistics. 2014;:550–9.
49. Keltner D, Sauter D, Tracy J, Cowen A. Emotional expression: advances in basic emotion theory. J Nonverbal Behav. 2019;43:133–60.

Publisher's note