Journal of Safety Science and Resilience 2 (2021) 246–252

Contents lists available at ScienceDirect

Journal of Safety Science and Resilience

journal homepage: http://www.keaipublishing.com/en/journals/journal-of-safety-science-and-resilience/

Improving sentiment analysis accuracy with emoji embedding

Chuchu Liu[a], Fan Fang[a], Xu Lin[b], Tie Cai[c], Xu Tan[c], Jianguo Liu[d], Xin Lu[a,*]

[a] *College of Systems Engineering, National University of Defense Technology, Changsha, 410073, China*
[b] *College of Computer, National University of Defense Technology, Changsha, 410073, China*
[c] *School of Software Engineering, Shenzhen Institute of Information Technology, Shenzhen, 518172, China*
[d] *Institute of Accounting and Finance, Shanghai University of Finance and Economics, Shanghai, 200433, China*

ARTICLE INFO

Keywords:
Sentiment analysis
Emoji
CEmo-LSTM
Sentiment evolution
COVID-19

ABSTRACT

Due to the diversity and variability of Chinese syntax and semantics, accurately identifying and distinguishing individual emotions from online texts is challenging. To overcome this limitation, we incorporate a new source of individual sentiment, emojis, which contain thousands of graphic symbols and are increasingly being used for expressing emotion in online conversations. We examined popular sentiment analysis algorithms, including rule-based and classification algorithms, to evaluate the impact of supplementing emojis as additional features to improve the algorithm performance. Emojis were also translated into corresponding sentiment words when constructing features for comparison with those directly generated from emoji label words. In addition, considering different functions of emojis in texts, we classified all posts in the dataset by their emoji usage and examined the changes in algorithm performance. We found that emojis are effective as expanding features for improving the accuracy of sentiment analysis algorithms, and the algorithm performance can be further increased by taking different emoji usages into consideration. In this study, we developed an improved emoji-embedding model based on Bi-LSTM (namely, CEmo-LSTM), which achieves the highest accuracy (around 0.95) when analyzing online Chinese texts. We applied the CEmo-LSTM algorithm to a large dataset collected from Weibo from December 1, 2019 to March 20, 2020 to understand the sentiment evolution of online users during the COVID-19 pandemic. We found that the pandemic remarkably impacted individual sentiments and caused more passive emotions (e.g., horror and sadness). Our novel emoji-embedding algorithm creatively combined emojis as well as emoji usage with the sentiment analysis model and can handle emotion mining tasks more effectively and efficiently.

1. Introduction

Sentiment analysis (SA) [1,2] aims to extract and identify the affective states or subjective opinions from texts and is generally considered to be a natural language processing (NLP) technique. A basic task in SA is classifying the polarity of a given text and determining whether the expressed emotion in the sentence (or document) is positive, negative, or neutral. With the advancement of machine learning and deep learning, different kinds of classification algorithms are widely used in SA tasks. Since users' posts on social networking platforms are generally short and concise, lexicon-based approaches to SA are also used frequently. However, although various advanced SA methods have been proposed, accurately identifying and classifying personal emotions from online texts is still challenging due to the diversity and variability of Chinese syntax and semantics. In addition, the rapid changes in internet slang further intensify the difficulty of understanding Chinese texts.

In recent years, users on social networks have become accustomed to utilizing a set of graphic symbols in online conversations to express their emotions. These emotional characters are called emojis [3] and have become a universal language used across apps, platforms, and cultures [4–7]. In general, the function of emojis is to provide emotional cues otherwise missing from typed conversation. Compared with words, emojis can express individual emotions more accurately and directly. When analyzing short informal texts, such as tweets, blogs, or comments, emojis provide crucial information regarding user sentiment [8–13]. The prevalence of emojis provides us with a new opportunity to analyze sentiment expressions in textual contexts.

Recent work in SA attempted to employ emotional emojis as noisy labels of sentiments on social media [14, 15]. In addition, emoji-labeled texts were used to train classifiers [16]. For example, Go et al. [17] identified the tweet polarity using emojis as noisy labels and collected a training dataset of 1.6 million tweets. Imtiaz et al. [18] directly used emojis as indicators of developers' sentiments on GitHub. However, the performance of such models can be limited due to noise in the labels. Recently, researchers started to analyze emojis from other perspectives to address the challenge of SA. Lu et al. [19] analyzed the use of emojis on GitHub and found that they are often used to express sentiment.

[*] Corresponding author
E-mail address: xin.lu@flowminder.org (X. Lu).

https://doi.org/10.1016/j.jnlssr.2021.10.003
Received 3 September 2021; Received in revised form 8 October 2021; Accepted 10 October 2021

Table 1
The basic information of the annotated corpus

Sentiment polarity	***Positive***	***Negative***	***Neutral***
Number of posts	4926	2752	2322
Labeling in dataset	1	-1	0
Emoji usage	***Strengthening***	***Reversing/ Revising***	***Uncertain***
Number of posts	4090	1048	361
Labeling in dataset	1	-1	0

Table 2
Example of emoji usage in Weibo posts.

Weibo post	**Emoji usage**
The medical staff have worked so hard. Thanks!!! 🥰	1
Awesome! Hope my idol make another great achievement. 🤗 😍	1
So we should worship and admire you, right?? 😏 🤮 🤮	-1
Without a mask? You are so brave. 😒 😒	-1

Mohammad et al. [20] constructed SVM classifiers with sparse indicator features, including n-grams, POS tags, punctuation, and emojis. Calefato et al. [21] and Ding et al. [22] took emoticons into account in their proposed SA techniques. All of them demonstrated the feasibility of leveraging these emotional cues to benefit SA. However, these studies mainly considered emojis as one feature and did not research the sentiment effects of emojis on the whole texts. Little attention has been given to the SA model combined with different emoji usages in texts.

In this study, we proposed an emoji-embedding architecture named CEmo-LSTM to improve the accuracy of sentiment identification and classification in SA tasks. We further evaluated the benefits of introducing emojis to the accuracy of SA in both the traditional rule-based and supervised learning algorithms. Additionally, the most effective approach for embedding emojis in SA algorithms was examined. We compared the performance of the CEmo-LSTM model with that of other mainstream SA models in different experimental settings. Finally, by collecting all posts and embedded emojis published by users on Weibo during the COVID-19 outbreak, we utilized CEmo-LSTM to analyze the sentiment evolution of online users and measured the impact of the COVID-19 pandemic on individual moods. To the best of our knowledge, this is the first study that comprehensively evaluates the effectiveness of introducing emoji usage into SA algorithms.

2. Data and method

2.1. Data collection

Weibo is a popular Twitter-like social media platform in China, which provides a rich publicly available data source for opinion mining and SA. We collected all data from Weibo that were posted publicly by users located in Wuhan (the capital of the Hubei province in China), including microblog text, posting time, author ID, and gender, from December 1, 2019 to March 20, 2020. By comparing the sentiments in posts published by Wuhan users before and after the COVID-19 outbreak, we can analyze the sentiment evolution of online users and further explore the impact of COVID-19 on individual moods. Overall, 38,183,194 microblog posts from 2,239,472 unique users were collected. We found that emotion tokens (i.e., emoji characters) were commonly used in Weibo posts. There were 15,609,843 posts containing emoji symbols, accounting for 40.88% of the total posts. In addition, 1,279,828 users used emojis at least once, accounting for 57.15% of all unique users.

2.2. Annotation

Although there have been some annotated corpora on Chinese and English for SA [23,24], they do not explicitly model the interaction between emojis and text. To fill in this gap, we manually annotated a Chinese microblog corpus. A total of 10 annotators (graduate students majoring in data analytics) were engaged to label the corpus, which consists of 10,000 randomly selected microblog posts. The sentiment polarities of the posts were manually classified as *positive, negative*, and *neutral*, denoted by 1, -1, and 0, respectively (Table 1). The annotators were asked to label each post by considering both the plain text and embedded emojis.

As there are several principal functions for which emojis are used (e.g., sentiment expression, sentiment enhancement, and sentiment modification) [25], the emoji usage of each post containing emojis was also annotated. Specifically, the emoji usage of each post was classified into three categories, *strengthening, reversing* (or *revising*), and *uncertain*, labelled by 1, -1, and 0, respectively, indicating whether the sentiment of the embedded emojis was consistent (1) or inconsistent (-1) with the sentiment of the text-only post (Table 2). The label 0 was used to denote when the effect of emojis in the post could not be confidently determined. We found that most emojis embedded in the posts were used to strengthen and clarify the sentiment of the original texts, accounting for approximately 73.6% of all posts with emojis included in the corpus. Finally, all 10,000 microblog posts were labelled with their sentiment polarities, of which 5499 posts containing emojis were also annotated with their emoji usages.

2.3. CEmo-LSTM model

In this study, we proposed a deep learning architecture, named the Chinese emoji-embedding LSTM model (CEmo-LSTM), to exploit the impact of emojis on sentiment analysis. Specifically, CEmo-LSTM introduces the emoji usage in online posts based on bidirectional long short-term memory (Bi-LSTM) [26]. Both plain texts and embedded emojis are used as input features, but before feature construction, the training corpus needs to be marked and filtered by different emoji usages. Since annotators can denote the emoji usage of each post when labeling the sentiment polarity of the training dataset, the workload for data annotation is not greatly increased.

As illustrated in Figure 1, our model includes the input sentence, word (emoji) representation, word embedding layer, Bi-LSTM layer, dropout layer, and a softmax layer. Given an input post S_i, the model first classifies the post according to whether there are any emojis embedded and evaluates the emoji usage of each post containing emojis. For posts containing emojis, both texts and emojis are input as features. Then, a microblog post can be described as $\{w_1, w_2, \ldots, w_i; E\}$, where w_i denotes the word token and E denotes the emoji. Through the embedding layer, both w_i and E are converted to the vector representation, d_i, as the input of the deep learning model to predict the sentiment polarity of a post. A Bi-LSTM layer is built to capture the representation of a microblog post, and a dropout layer is added to prevent over-fitting and improve the generalizability of the model. Finally, a softmax activation function is used to calculate a probability distribution p over a set of sentiment polarities $\{1, -1, 0\}$. Consequently, a list of labels of input posts is predicted according to the corresponding output of the softmax layer.

3. Experiments

3.1. Experimental setting

To evaluate the performance of CEmo-LSTM, we have to prove the impact of emojis on the sentiment identification of texts and discover the best approach for exploiting the novel emotional clue (i.e., emojis) in online posts for SA tasks. Specifically, our goal is to answer three research questions:

RQ1: Does the supplementation of emojis promote the emotion recognition of texts? To answer this question, a rigorous con-

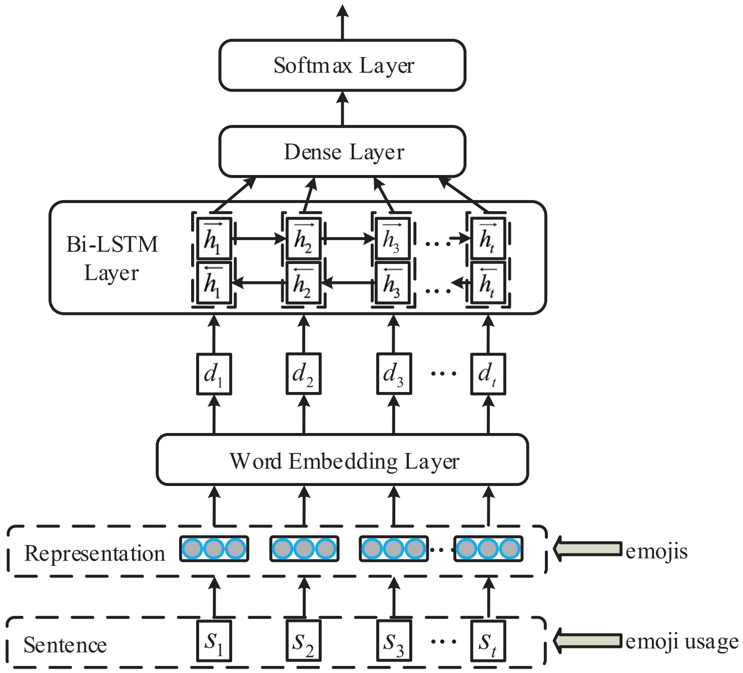

Softmax Layer
Dense Layer
Bi-LSTM Layer
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_3}$ … $\overrightarrow{h_t}$
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_3}$ … $\overleftarrow{h_t}$
d_1 d_2 d_3 … d_t
Word Embedding Layer
Representation
emojis
Sentence
S_1 S_2 S_3 … S_t
emoji usage

Figure 1. The architecture of the CEmo-LSTM model.

trast experiment was conducted. We compared the performance of SA algorithms on posts with embedded emojis and text-only posts, respectively, to measure the impact of emojis on emotion recognition. For text-only posts, a microblog text was described as $\{w_1, w_2, \ldots, w_i\}$, where w_i denotes the word token. A post with embedded emojis was represented as $\{w_1, w_2, \ldots, w_i ; E\}$, where E indicates the set of emoji tag words.

RQ2: Can the tag words of emojis be directly used when constructing features? We examined whether the vagueness and ambiguity of emoji tag words would affect the sentiment identification of SA algorithms. Before constructing features, all emojis were converted into corresponding sentiment words (e.g., Sad, Happy) instead of emoji tag words based on their meanings and sentiments, and we evaluated the changes in algorithm performance. Accordingly, an emoji-embedded post was denoted as $\{w_1, w_2, \ldots, w_i ; ES\}$, where ES is the set of sentiment words translated from emojis.

RQ3: Does the classification of the training dataset on emoji usage improve the performance of SA algorithms? Corresponding to this question, an experiment was also conducted. We classified the emoji usage of all posts containing emojis to examine the impact of the introduction of emoji usage on SA algorithms. We found that, in most posts on Weibo, the emotions expressed by emojis were consistent with emotions of plain texts, and the main function of emojis was to clarify and enhance the sentiment of the sentence. Hence, *strengthening* posts in the corpora (labelled with 1 in the field *emoji usage*) were filtered out and used to train SA models. A post classified by emoji usage U was described as $\{w_{u1}, w_{u2}, \ldots, w_{ui} ; EU\}$, where w_{ui} denotes the word token and EU stands for the set of emojis embedded.

In all experiments, text segmentation work was carried out with Jieba [27], a popular Chinese word segmentation package. We filtered out stop words, punctuation, and spaces in the posts to clean the data. For the CEmo-LSTM model, we set the dimensions of word and emoji embeddings as 200, and the dimensions of hidden and cell states in Bi-LSTM cells as 64. The dropout probability (the loss of some units at random during training) in the dropout layer was set to 0.4. The “categorical_crossentropy” loss function was chosen, and the RMSprop [28] method was used for optimizing the objective functions. Since CEmo-LSTM is an emoji-embedding SA model based on emoji usage, in the first two experiments for RQ1 and RQ2, CEmo-LSTM (text) represents CEmo-LSTM’s implementation in plain texts, CEmo-LSTM (text+E) represents the implementation with emoji tags embedded, and CEmo-LSTM (text+ES) denotes when emoji tags were replaced with sentiment words.

3.2. Baselines

To evaluate the effectiveness of the improved emoji-embedding SA model (CEmo-LSTM), we introduced several baseline models, including the state-of-the-art method, for comparison. Using the same experimental settings, we evaluated the performance of different algorithms. Both supervised and unsupervised learning methods were carried out.

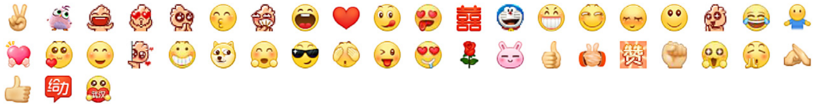
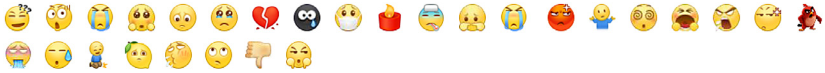
3.2.1. Rule-based approach

In general, the implementation of rule-based SA relies on a specific sentiment lexicon. In this study, we constructed two lexicons: the traditional lexicon for sentiment words (sentiment lexicon, for short) and an emoji lexicon based on the sentiment of different emojis. Based on these two lexicons, we extracted all sentiment words and emojis contained in each post. By measuring the frequency and emotional intensity of sentiment words (or emojis), each post was assigned a sentiment score. If the score was greater than 0, the post was considered positive. Finally, the accuracy of the algorithm was validated by comparing the results with the manual annotations of posts.

(1) Sentiment lexicon. To construct the sentiment lexicon, we first integrated four popular Chinese sentiment dictionaries, including DUTIR, C-LIWC, HowNet, and NTUSD [29,30]. Then, by supplementing popular sentiment words used on the internet [31], we built a comprehensive sentiment lexicon, which is more suitable for SA on Weibo.

(2) Emoji lexicon. As there is significant heterogeneity [32,33] in the popularity of different emojis (i.e., in the Sina Weibo data used), the top 100 most popular emojis account for approximately 96% of all emojis used daily. We constructed an emoji lexicon (Table 3) based on the top 100 most frequently used emojis and classified them into three categories, *positive, negative*, and *neutral*, according to their official annotations and emotions expressed. Each emoji was also

Table 3
Example of emoji lexicon.

Positive	
Negative	
Neutral	

assigned a sentiment value, with positive emojis denoted from 1 to 5 and negative emojis denoted from -1 to -5, respectively. The absolute value represents the emotional intensity.

3.2.2. Classification algorithms

A total of six mainstream classification algorithms, which are widely used in SA tasks due to their promising effectiveness, were evaluated. It is worth noting that the rule-based approach was only used to discover the impact of emojis on sentiment recognition (the first experiment). All six classification algorithms were implemented in all three experiments, and their performance was compared with that of the CEmo-LSTM model based on the same experimental settings. The detailed setting information for each algorithm is summarized below.

- Logistic Regression (LR) [34]: LR was carried out for SA tasks in all three experiments, with LR (text) representing the implementation of LR in plain texts, LR (text+E) representing the implementation in posts with emoji tags embedded, LR (text+ES) representing LR's operation when emoji tags were replaced by corresponding sentiment words, and LR (EU) representing LR's operation in posts classified by emoji usage. To train each model, the features of the posts were used as inputs, such as emojis, bag-of-words, and TF-IDF values.
- Support Vector Machine (SVM) [35]: SVM was also compared in each experiment, with SVM (text) denoting SVM's operation in plain texts, SVM (text+E) denoting that posts with emojis embedded, SVM (text+ES) denoting the introduction of emojis' sentiment words, and SVM (EU) relating to emoji usage. Similarly, emojis, bag-of-words, and TF-IDF values in the posts were used to train each classifier.
- Naive Bayes classifier (NB) [36]: For the NB algorithm, the Naive Bayes classifier for multinomial models (i.e., the multinomial Naive Bayes classifier) was used, which is suitable for classification with discrete features (e.g., word counts for text classification). Likewise, NB was carried out in each experiment, corresponding to NB (text), NB (text+E), NB(text+ES), and NB (EU). All parameters were kept the same in each NB operation.
- Gradient Boosting Decision Tree (GBDT) [37]: GBDT, the gradient boosting classifier, was also operated as GBDT (text), GBDT (text+E), GBDT (text+ES), and GBDT (EU). In each experiment, text features (or with emojis) were used to train the GBDT classifier. The learning rate was set to 0.05, and the number of estimators was 540.
- Long Short-Term Memory (LSTM) [38]: The experimental setups of LSTM were exactly the same as for CEmo-LSTM. The LSTM experiments were named LSTM (text), LSTM (text+E), LSTM (text+ES), and LSTM (EU), respectively, and each model learned the vector representations of words or emojis from microblog posts.
- Bidirectional Encoder Representation from Transformers (BERT) [39]: Similarly, BERT was carried out in all experiments, corresponding to BERT (text), BERT (text+E), BERT (text+ES), and BERT (EU). The dropout probability was set as 0.15, and Adam was used as the optimization method during training.

3.2.3. Evaluation metric

We used tenfold cross validation in our experiments. The original dataset was randomly split into ten equal sections. In each fold, nine sections were selected for training, and the tenth section was used for testing. The classification results were measured by accuracy, λ, which is the ratio of correctly identified sentiments of posts among all corpora, and defined as,

$$\lambda = {}^{T}/_{N}$$

where T indicates the number of predicted sentiment ratings that are identical with manual sentiment ratings, and N indicates the number of posts. In each experiment, we compared the accuracy of the CEmo-LSTM model with that of all baseline algorithms.

Table 4
Performance comparison of classification algorithms with (without) emojis as features.

Model	λ	Model	λ
LR (text)	0.692	LR (text+E)	**0.727**
SVM (text)	0.658	SVM (text+E)	**0.681**
NB (text)	0.692	NB (text+E)	**0.709**
GBDT (text)	0.659	GBDT (text+E)	**0.720**
LSTM (text)	0.733	LSTM (text+E)	**0.790**
BERT (text)	0.718	BERT (text+E)	**0.762**
CEmo-LSTM (text)	0.743	CEmo-LSTM (text+E)	**0.811**

3.3. Results

3.3.1. Effect of emojis on the accuracy of sentiment recognition

(1) **Rule-based approach.** To discover the effect of emojis on sentiment recognition, the classical rule-based approach for unsupervised learning was examined. We conducted the ruled-based algorithm both in posts with embedded emojis (emoji posts) and posts consisting plain texts (emoji-free posts). We found that the performance of the algorithm with emoji posts ($\lambda = 0.561$) was significantly better than with emoji-free posts ($\lambda = 0.360$). Emojis are beneficial clues for the rule-based algorithm in SA tasks. This further indicates that emojis play an important role in clarifying and enhancing the sentiment of sentences. However, the accuracy of the rule-based algorithm in both scenarios was not satisfactory, possibly due to the short length of internet micro-texts and inadequate emotional clues.

(2) **Classification algorithms.** In order to further evaluate the impact of emojis on sentiment recognition, the performance of the classification algorithms in supervised learning was examined in the two scenarios. As shown in Table 4, in most cases the accuracy of the classification algorithms with emoji posts was significantly higher than with emoji-free posts, indicating that in supervised learning algorithms the supplementation of emojis helps to clarify sentence emotions. In addition, we found that in sentiment classification of online micro-texts, algorithms using deep learning outperformed other classification algorithms. CEmo-LSTM followed by LSTM showed the highest accuracy with our dataset, which is also popularly applied for other text-based SA tasks [40]. Generally, classification algorithms were superior to the rule-based (unsupervised) algorithm in accuracy. Because the operation of unsupervised learning does not rely on manual annotations and the sentiment lexicon can be updated based on

Table 5
Converting emojis into corresponding emotional words.

Model	λ	Model	λ
LR (text+E)	0.727	LR (text+ES)	0.727
SVM (text+E)	**0.681**	SVM (text+ES)	0.628
NB (text+E)	**0.709**	NB (text+ES)	0.691
GBDT (text+E)	0.720	GBDT (text+ES)	0.720
LSTM (text+E)	**0.790**	LSTM (text+ES)	0.766
BERT (text+E)	**0.762**	BERT (text+ES)	0.735
CEmo-LSTM (text+E)	**0.811**	CEmo-LSTM (text+ES)	0.774

Table 6
The performance of different SA algorithms based on emoji usage.

Model	λ	Model	λ
LR (text+E)	0.727	LR (**EU**)	**0.750**
SVM (text+E)	0.681	SVM (**EU**)	**0.792**
NB (text+E)	0.709	NB (**EU**)	**0.813**
GBDT (text+E)	0.720	GBDT (EU)	0.720
LSTM (text+E)	0.790	LSTM (**EU**)	**0.939**
BERT (text+E)	0.762	BERT (**EU**)	**0.853**
CEmo-LSTM (text+E)	0.811	**CEmo-LSTM**	**0.949**

specific datasets, the rule-based SA algorithm is also frequently used in practical scenarios.

3.3.2. *Feature comparison between emoji tag words and sentiment words*

It was assumed that the ambiguity of emoji tag words would affect the understanding of sentence emotions for SA algorithms. Consequently, all emoji tags in posts were replaced with corresponding sentiment words when constructing features. However, the empirical results fail to prove this hypothesis, and the accuracy of all algorithms unexpectedly decreased (Table 5). Replacing tag words of emojis with sentiment words slightly reduced the performance of the original algorithms. This indicates that the ambiguity of emoji tags has no negative impact on sentiment classification in practice, and they can be used as effective features in SA tasks.

3.3.3. *Improving algorithm accuracy with sentiment strengthening*

To evaluate the impact of the introduction of emoji usage in SA algorithms, *strengthening* posts in the corpora were filtered out and used to train SA models. We found that the accuracy of each classification algorithm significantly improved after examining the consistency between the sentiments of emojis and those of plain texts. This indicates that posts in which the emoji sentiment is inconsistent with the text sentiment tend to reduce the performance of SA algorithms. Before training SA models, it is useful to classify the training dataset with emoji usage. This also proves the rationality of the architecture design of the CEmo-LSTM model. As shown in Table 6, although the introduction of emoji usage dramatically improved the accuracy of all algorithms, our improved emoji-embedding model (CEmo-LSTM) always provided the best performance in SA tasks.

4. Case study

As the COVID-19 pandemic sweeps across the world, it is causing widespread concern, fear, and stress. Some studies indicated that the pandemic not only threatened physical health but also affected individual mentality and emotions [41]. To understand the potential emotional changes of Wuhan residents caused by COVID-19, we collected all posts published by Weibo users who were located in Wuhan during the COVID-19 outbreak and conducted SA on the dataset utilizing the CEmo-LSTM algorithm.

The percentage of positive posts and the percentage of negative posts published daily were calculated, respectively. We found that after the COVID-19 outbreak the number of positive posts on Weibo dropped drastically (Figure 2A). This result verifies the above conclusion that the pandemic has had psychologically negative effects on individuals. Furthermore, in order to examine the evolution of specific sentiments of Wuhan residents, we divided user sentiments into seven categories: *Happy, Appreciated, Angry, Sad*, *Scared, Disgusted*, and *Surprised* [42]. The evolution pattern of each sentiment was analyzed (Figure 2B). It can be seen that after the outbreak of COVID-19, with the spread of the novel coronavirus, the proportion of posts related to *Sad* and *Scared* made by Wuhan users clearly increased. By further mining the textual content of these posts, we found that most topics were relevant to the spread, treatment, and impact of COVID-19. In general, the outbreak of the pandemic has indeed caused more negative emotions for Wuhan residents.

5. Conclusion & discussion

Due to the diversity of Chinese expressions and the variability of Chinese syntax and semantics, SA algorithms are unable to achieve satisfactory results when processing Chinese texts, especially short micro-texts. Emojis, which are graphic symbols carrying specific meanings, have been frequently embedded within micro-texts to more directly express emotional meanings, and they provide novel information on user sentiments. As emojis are widely adopted in online conversations across

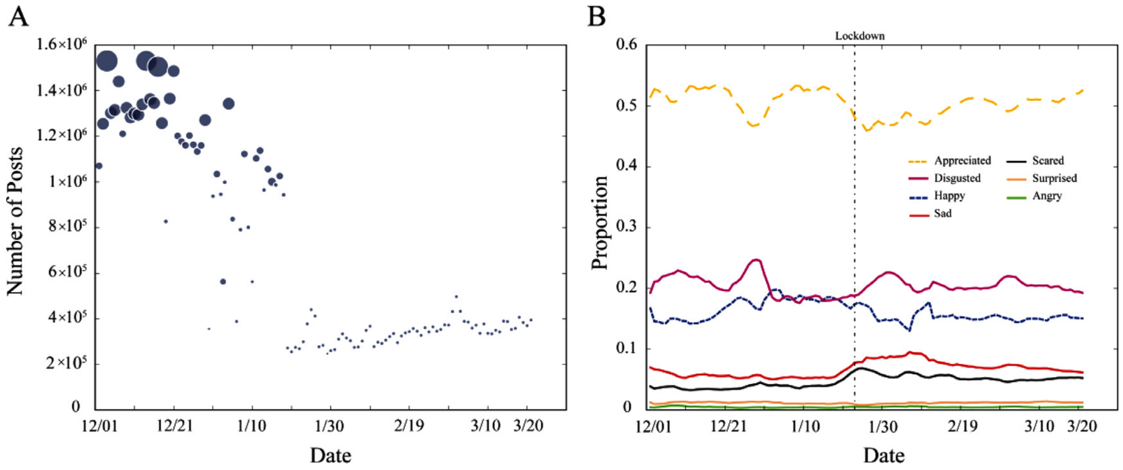


Figure 2. (A) Evolution of positive posts; (B) Evolution of the proportion of different sentiments.

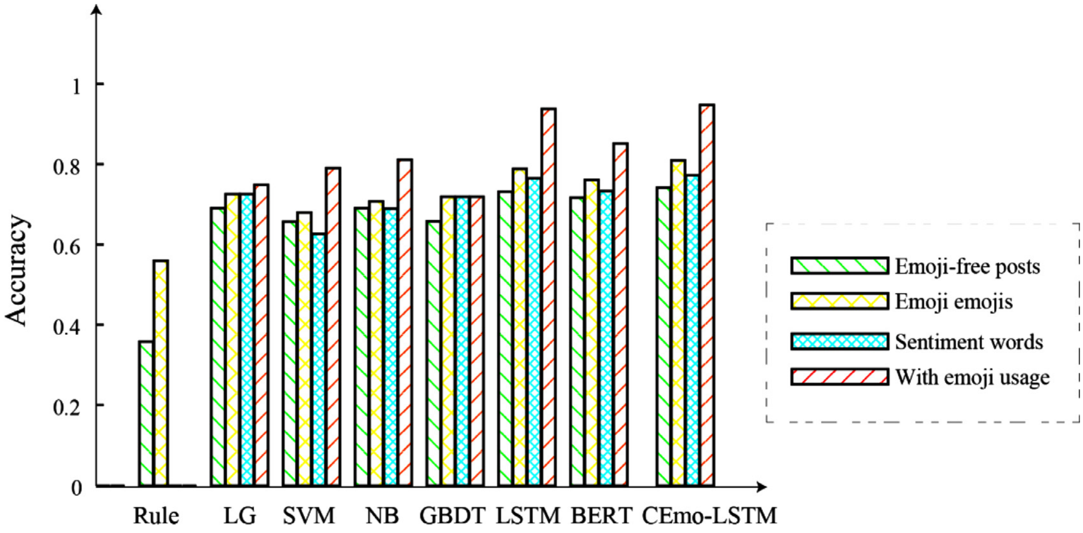


Figure 3. The performance of all SA algorithms.

apps and platforms, they can be introduced into SA algorithms as crucial features to improve performance.

In this study, we examined and compared popular SA algorithms, including the rule-based algorithm in unsupervised learning, classification algorithms (e.g., SVM), and neural network algorithms (e.g., LSTM) in supervised learning. The effect of emoji introduction and the ambiguity of emoji tags were also evaluated. We found that the accuracy of supervised learning algorithms is generally higher than that of unsupervised learning algorithms. Further, deep learning algorithms (e.g., LSTM and Bi-LSTM) always achieve the best performance. In addition, we found that introducing emojis is beneficial to improve the performance of SA algorithms, and emoji tag words can be used directly when constructing features for classifier training. It is worth noting that after classifying the emoji usage of posts in the training set, the performance of each algorithm improved significantly. An overview of all algorithms and their improvements is shown in Figure 3.

Accordingly, combined with emoji usage, we developed an improved emoji-embedding model based on Bi-LSTM (namely, CEmo-LSTM), in which emojis are used as one of the features, and the training data are classified by their emoji usage before training the classifier. Compared with existing SA algorithms [43], our model achieved the highest accuracy when analyzing online Chinese texts. Finally, the proposed algorithm, CEmo-LSTM, was applied to the SA of Wuhan residents during the COVID-19 outbreak. It was found that the pandemic has had a negative impact on individual sentiments, and the outbreak has resulted in more passive emotions (e.g., scared and sad) on the part of Wuhan residents.

This study proposed a novel emoji-embedding algorithm based on emoji usage for SA, highlighting the sentiment evolution of social platform users due to the COVID-19 outbreak. However, our focused emotional symbols in this study were mainly from the common emojis of Sina, and there are many different emoji packs from other sources, it is necessary to explore the usage patterns of more massive emojis and to implement comparative studies across platforms and contexts. In addition, different usage habits of emojis may lead to different sentiment semantics. We tend to analyze various emoji use habits of online users in detail to promote the comprehension of emotional meanings conveyed by emojis, and explore more complex contexts to further improve the performance of the CEmo-LSTM algorithm in future work.

Declaration of Interests

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

Acknowledgements

CL and FF were partially supported by the National Natural Science Foundation of China (82041020, 72088101, 91846301). XL acknowledges support from the National Natural Science Foundation of China (72025405, 71771213) and the Hunan Science and Technology Plan Project (2020JJ4673, 2020TP1013). JL was supported by the National Natural Science Foundation of China (61773248) and the Major Program of National Fund of Philosophy and Social Science of China (20ZDA060). TC and XT were supported by the Shenzhen Basic Research Project for Development of Science and Technology (JCYJ20200109141218676).

References

[1] R. Feldman, Techniques and applications for sentiment analysis, Commun. ACM 56 (4) (2013) 82–89.

[2] W Medhat, A Hassan, H. Korashy, Sentiment analysis algorithms and applications: a survey], Ain Shams Eng. J. 5 (4) (2014) 1093–1113.

[3] Description of emoji. [cited 2021 20 May]. Available from: https://en.wikipedia.org/wiki/Emoji

[4] N Ljubešić, D. Fišer, A global analysis of emoji usage, in: Proceedings of the 10th Web as Corpus Workshop, 2016, pp. 82–89.

[5] W Ai, X Lu, X Liu, et al., Untangling emoji popularity through semantic embeddings, Eleventh International AAAI Conference on Web and Social Media, 2017.

[6] P K Novak, J Smailović, B Sluban, et al., Sentiment of emojis, PLoS One 10 (12) (2015).

[7] M Shiha, S. Ayvaz, The effects of emoji in sentiment analysis, Int. J. Comput. Electr. Eng.(IJCEE.) 9 (1) (2017) 360–369.

[8] M Boia, B Faltings, CC Musat, P. Pu, A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets, Intl. Conf. Soc. Comput. (SocialCom). (2013) 345–350.

[9] A Hogenboom, D Bal, F Frasincar, M Bal, Jong F de, U. Kaymak, Exploiting emoticons in polarity classi-fication of text, J. Web Eng. 14 (1–2) (2015) 22–40.

[10] D Davidov, O Tsur, A. Rappoport, Enhanced sentiment learning using Twitter hashtags and smileys, in: Proc. 23rd Intl. Conf. on Computational Linguistics: Posters., ACL, 2010, pp. 241–249.

[11] KL Liu, WJ Li, M. Guo, Emoticon smoothed language models for Twitter sentiment analysis, in: Proc. 26th AAAI Conf. on Artificial Intelligence, 2012, pp. 1678–1684.

[12] J. Read, Using emoticons to reduce dependency in machine learning techniques for sentiment classification, in: Proc. ACL Student Research Workshop, ACL, 2005, pp. 43–48.

[13] J Zhao, L Dong, J Wu, K. Xu, Moodlens: An emoticon-based sentiment analysis system for Chinese tweets, in: Proc. 18th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1528–1531.

[14] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, Xuanzhe Liu, Emoji-powered representation learning for cross-lingual sentiment classification, in: Proceedings of the 2019 World Wide Web Conference on World Wide Web, WWW 2019, 2019, pp. 251–262.

[15] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, Sune Lehmann, Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, 2017, pp. 1615–1625.

[16] Sascha Narr, Michael Hulfenhaus, Sahin Albayrak, Language-independent Twitter sentiment analysis, in: Proceedings of the Learning, Knowledge, and Adaption Conference (LWA'12), 2012, pp. 12–14.

[17] Alec Go, Richa Bhayani, Lei Huang, Twitter Sentiment Classification Using Distant Supervision, CS224N Project Report, Stanford University, 2009.

[18] Nasif Imtiaz, Justin Middleton, Joymallya Chakraborty, Neill Robson, Gina Bai, Emerson R. Murphy-Hill, Investigating the effects of gender bias on GitHub, in: Proceedings of the 41st International Conference on Software Engineering, ICSE 2019, 2019, pp. 700–711.

[19] Xuan Lu, Yanbin Cao, Zhenpeng Chen, Xuanzhe Liu, A first look at emoji usage on GitHub: an empirical study, CoRR abs/1812 2018 (2018) 04863.

[20] Saif M. Mohammad, Svetlana Kiritchenko, Xiaodan Zhu, NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets, in: Proceedings of the Joint Conference on Lexical and Computational Semantics, 2013, pp. 321–327.

[21] Fabio Calefato, Filippo Lanubile, Federico Maiorano, Nicole Novielli, Sentiment polarity detection for software development, Empirical Software Eng. 23 3 (2018) (2018) 1352–1382.

[22] Jin Ding, Hailong Sun, Xu Wang, Xudong Liu, Entity-level sentiment analysis of issue comments, in: Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering, SEmotion@ICSE 2018, 2018, pp. 7–13.

[23] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, Veselin Stoyanov, SemEval-2016 Task 4: Sentiment analysis in Twitter, in: Proceedings of the International Workshop on Semantic Evaluation, 2016, pp. 1–18.

[24] Debora Nozza, Elisabetta Fersini, Enza Messina, A multi-view sentiment corpus, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 1, Long Papers, 2017, pp. 273–280.

[25] Guibon G, Ochs M, Bellot P. From emojis to sentiment analysis//WACAI 2016. 2016.

[26] B Jang, M Kim, G Harerimana, et al., Bi-LSTM model to increase accuracy in text classification: combining Word2vec CNN and attention mechanism, Appl. Sci. 10 (17) (2020) 5841.

[27] Project description of jieba. [cited 2021 20 May]. Available from: https://pypi.org/project/jieba/.

[28] F Zou, L Shen, Z Jie, et al., A sufficient condition for convergences of adam and rmsprop//, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11127–11135.

[29] X Fu, G Liu, Y Guo, Z. Wang, Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, Knowl.-Based Syst. 37 (2013) 186–195.

[30] S Liu, J. Chen, A multi-label classification based approach for sentiment classification, Expert Syst. Appl. 42 (3) (2015) 1083–1093.

[31] Y M Zhou, A M Yang, J H Lin, A method of building Chinese microblog sentiment lexicon, J. Shandong Univ. (Engineering Science) 44 (2014) 36–40.

[32] MEJ. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. 46 (5) (2005) 323–351.

[33] S Aoki, M. Nirei, Zipf's law, Pareto's law, and the evolution of top incomes in the United States, Am. Econ. J.: Macroeconomics 9 (3) (2017) 36–71.

[34] Y Kanbayashi, K Sakaguchi, F Hongo, et al., Predictors for development of denosumab-induced hypocalcaemia in cancer patients with bone metastases determined by ordered logistic regression analysis, Sci. Rep. 11 (1) (2021) 1–6.

[35] W S Noble, What is a support vector machine? Nat. Biotechnol. 24 (12) (2006) 1565–1567.

[36] I. Rish, An empirical study of the naive Bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, 3, 2001, pp. 41–46.

[37] H Rao, X Shi, A K Rodrigue, et al., Feature selection based on artificial bee colony and gradient boosting decision tree, Appl. Soft Comput. 74 (2019) 634–642.

[38] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling//Thirteenth annual conference of the international speech communication association. 2012.

[39] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[40] A Yadav, D K Vishwakarma, Sentiment analysis using deep learning architectures: a review, Artif. Intell. Rev. 53 (6) (2020) 4335–4385.

[41] S Li, Y Wang, J Xue, et al., The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users, Int. J. Environ. Res. Public Health 17 (6) (2020) 2032.

[42] Chen Jianmei, The Construction and Application of Chinese Emotion Word Ontology, Dalian: Dalian University of Technology, 2009.

[43] Y Lou, Y Zhang, F Li, et al., Emoji-Based Sentiment Analysis Using Attention Networks, ACM Trans. Asian and Low-Resource Language Inf. Process. (TALLIP) 19 (5) (2020) 1–13.