

A Novel Machine Learning-based Sentiment Analysis Method for Chinese Social Media Considering Chinese Slang Lexicon and Emoticons

Da Li¹, Rafal Rzepka¹, Michal Ptaszynski²,
and Kenji Araki¹

¹ Graduate School of Information Science and Technology
Hokkaido University, Sapporo, Japan

² Department of Computer Science, Kitami Institute of Technology, Kitami, Japan

Abstract. Internet slang is an informal language used in everyday online communication which quickly becomes adopted or discarded by new generations. Similarly, pictograms (emoticons/emojis) have been widely used in social media as a mean for graphical expression of emotions. People can convey delicate nuances through textual information when supported with emoticons. Furthermore, we also noticed that when people use new words and pictograms, they tend to express a kind of humorous emotion which is difficult to clearly classify as positive or negative. Therefore, it is important to fully understand the influence of Internet slang and emoticons on social media. In this paper, we propose a machine learning method considering Internet slang and emoticons for sentiment analysis of Weibo, the most popular Chinese social media platform. In the first step, we collected 448 frequent Internet slang expressions as a slang lexicon, then we converted the 109 Weibo emoticons into textual features creating Chinese emoticon lexicon. To test the capability of recognizing humorous posts, we utilized both lexicons with several machine learning approaches, k-Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, Naïve Bayes and Support Vector Machine for detecting humorous expressions on Chinese social media. Our experimental results show that the proposed method can significantly improve the performance for detecting expressions which are difficult to polarize into positive-negative categories.

Keywords: sentiment analysis · machine learning · social media · Internet slang · emoticons

1 Introduction

Nowadays, people have become increasingly accustomed to expressing their opinions online, especially on social media such as Twitter, Facebook or Weibo - the biggest Chinese social media network that was launched in 2009. The rapid growth of such platforms provides rich multimedia data in large quantities for various research opportunities as sentiment analysis which focuses on automatic sentiment prediction on given contents. Microblog data contain a vast amount of valuable sentiment information not only for the commercial use, but also for psychology, cognitive linguistics or political

science. Sentiment analysis has been widely used in real world applications by analyzing the online user-generated data, such as election prediction, opinion mining and business-related activity analysis [25]. Sentiment analysis of microblogs becomes an important area of research in the field of Natural Language Processing. Study of sentiment in microblogs in English language has undergone major developments in recent years [14]. Chinese sentiment analysis research, on the other hand, is still at relatively early stage [19] especially when it comes to lexicons and emoticons usage.

Pictograms (emoticons/emojis) have been widely used in social media as a mean for graphical expression of emotions. According to the study about Instagram, emojis are present in up to 57% of online messages in many countries³. For example, 😄 “face with tears of joy”, an emoji that means that somebody is in an extremely good mood, was regarded as the 2015 word of the year by The Oxford Dictionary [12]. In our opinion ignoring emoticons in sentiment research is unjustifiable, because they convey a significant emotional information and play an important role in expressing emotions and opinions in social media [13, 5].

Internet slang is ubiquitous on the Internet. The emergence of new social contexts like micro-blogs, question-answering forums, and social networks has enabled slang and non-standard expressions to abound on the web. Despite this, slang has been traditionally viewed as a form of non-standard language, a form of language that is not the focus of linguistic analysis and has largely been neglected [8].

Furthermore, we also noticed that when people use new words and pictograms, they tend to express a kind of humorous emotion which is difficult to be easily classified as positive or negative. It seems that some emoticons are used just for fun, self-mockery or jocosity which expresses an implicit humor which might be characteristic to Chinese culture. Emoticons and slang seem to play an important role in expressing this kind of emotion. There is a high possibility that this phenomenon can cause a significant difficulty in sentiment recognition task. Figure 1 shows an example of a Weibo microblog posted with emoticons and Internet slang. In the second line of the post, 累觉不爱 (*lei jue bu ai*) is a Chinese informal contraction meaning 很累, 感觉自己不会再爱了 (*hen lei, gan jue zi ji bu hui zai ai le* which means “too tired for romance”). Such abbreviations are popular and usually extracted from popular phrases and shorten into four characters in general, and become a new *chengyu*, a type of traditional Chinese idiomatic expression most of which consist of four characters. *Chengyu* are considered as collected wisdom of the Chinese culture. Through the insights learned from *chengyu*, we can express and discover wise men’s experiences, moral concepts, or admonishments from the older generations of Chinese. Nowadays, *chengyu* still plays an important role in Chinese conversations and education. When a new *chengyu* is introduced it can also convey a humorous content. Examples of such abbreviations are: *lei jue bu ai*, *ren jian bu chai* (life is so hard that some lies are better not exposed), *xi da pu ben* (news so exhilarating that everyone is celebrating and spreading it around the world) and so on. When it comes to emoticons, new ones are introduced by social media companies, but their meaning can change with time. For example 😊 was originally and emoji meant for expressing “bye-bye” gesture. However, it seems that gradually Weibo users

³ <https://www.quintly.com/blog/instagram-emoji-study>

started using this emoji for expressing artificial smile and refuse or self-mockery⁴. In the research of [9], it was shown that this emoji expresses humorous emotion rather than negative polarity. For example, in the following post: “After jogging, I’m starving. Someone sent me a picture of kebab. I’m too tired for romance 🙄🙄🙄”.



Fig. 1. Example of Weibo post with Internet slang and emoticons. The entry says “After jogging, I’m starving. Someone sent me a picture of kebab. I’m too tired for romance”.

To address this phenomenon, in this paper we focus on the Internet slang and emoticons used on Weibo in order to establish if both slang and emoticons improve sentiment

⁴ <https://qz.com/944693>

analysis by recognizing humorous entries which are difficult to polarize. To perform experiments, we collected 448 frequent Chinese Internet slang expressions as a slang lexicon, then we converted 109 Weibo emoticons into textual features creating Chinese emoticon lexicon. Then we utilized both lexicons with several machine learning approaches, k-Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, Naïve Bayes and Support Vector Machine for detecting humorous expressions on Chinese social media. Our experimental results show that the proposed method can significantly improve the performance for detecting expressions which are difficult to polarize into positive-negative categories.

Our main contributions are as follows:

- We collected 448 frequent Chinese Internet slang expressions as a Chinese slang lexicon.
- We converted the 109 Weibo emoticons into textual features creating Chinese emoticon lexicon.
- We empirically confirmed implicit humor characteristic to Chinese culture visible on Weibo and utilized both lexicons with several machine learning approaches for detecting humorous expressions on Weibo and confirmed that using both slang and emoticons improves previously proposed method.

2 Related Research

At present, the sentiment analysis technology generally can be divided into two categories: rule-based methods relying on sentiment lexicons, and machine learning-based methods relying on annotated data.

2.1 Rule-based methods

Zhang et al. [24] proposed a rule-based approach with two phases: a) the sentiment of each sentence is first decided based on word dependency to aggregate the sentences sentiments and then b) the sentiment of each document is calculated. Zagibalov et al. [23] presented a method that does not require any annotated corpus training data and only requires information on commonly occurring negations and adverbials. Li et al. stated that polarities and strengths judgment of sentiment words comply with a Gaussian distribution, and thus proposed a Normal distribution-based sentiment computation method which allows quantitative analysis of semantic fuzziness of sentiment words in Chinese language [10]. Zhuo et al. presented a novel approach based on the fuzzy semantic model by using an emotion degree lexicon and a fuzzy semantic model [26]. Their model includes text preprocessing, syntactic analysis, and emotion word processing. However, optimal results of Zhuo's model were achieved only when the task was clearly defined. Wu et al. presented an approach to leverage Web resources to construct a English Slang Sentiment Dictionary (SlangSD) that is easy to expand [21]. They empirically showed the advantages of using SlangSD, the newly-built slang sentiment word dictionary for sentiment classification, and provided examples demonstrating its ease of use with a sentiment analysis system.

2.2 Machine learning-based methods

Tan and Zhang conducted an empirical study of sentiment categorization on Chinese documents [18]. They tested four features – mutual information, information gain, chi-square, and document frequency; and five learning algorithms: centroid classifier, k-Nearest Neighbor, Winnow classifier, Naïve Bayes (NB) and Support Vector Machine (SVM). Their results showed that the information gain and SVM features provided the best performances for sentiment classification coupled with domain or topic dependent classifiers. There are also researchers who have combined the machine learning approach with the lexicon-based approach. Chen et al. proposed a novel sentiment classification method which incorporated existing Chinese sentiment lexicon and convolutional neural network [2]. The results showed that their approach outperforms the convolutional neural network (CNN) model only with word embedding features [7]. However, all these approaches did not consider emoticons.

Recently, a powerful system utilizing emoji in Twitter sentiment analysis model called DeepMoji was proposed [4]. Its creators trained 1,246 million tweets containing one of 64 common emoticons by Bi-directional Long Short-Term Memory (Bi-LSTM) model and applied it to interpret the meaning behind the online messages. DeepMoji is also the most advanced sarcasm-detecting model, with an accuracy rate of 82.4% even outperforming human detectors who managed to acquire 76.1% accuracy rate. Sarcasm reverses the emotion of the literal text, therefore sarcasm-detecting capability can play a significant role in sentiment analysis, especially in case of social media. Although sarcasm and irony tend to convey negative emotions in general, we found that in Chinese social media (Weibo in our example), in addition to the expression of positive and negative emotions, people tend to express a kind of humorous emotion that escapes the traditional bi-polarity.

Table 1. Examples of our Chinese Internet Slang Lexicon.

Type	Examples (Origin)	English Translation
Numbers	233 (哈哈)	“laughter”
Latin alphabet abbreviations	TMD (他妈的)	“Damn”
Chinese contractions	人艰不拆 (人生已经如此的艰难 有些事情就不要拆穿)	“Life is so hard that some lies are better not exposed.”
Neologisms	屌丝	“Loser”
Phrases with altered or extended meanings	壕 (土豪)	“Vulgar tycoon”
Puns and wordplay	河蟹 (和谐)	“Harmony”
Slang derived from foreign language	欧尼酱 (お兄ちゃん)	“Brother”

3 Lexicon of Chinese Online Slang

Chinese Internet slang is informal language used to express ideas on the Chinese Internet in response to events, to mass media and foreign cultures. It also expresses a natural human desire to simplify and update language. Slang that first appears on-line is often adopted to become widely used in everyday life. It includes content relating to all aspects of social life, mass media, economic, political situation etc. Internet slang is arguably the fastest-changing aspect of a language, created by a number of different influences, technology, mass media and foreign culture amongst others.

Because Internet slang is not easy to extract automatically, it can cause a significant difficulty in sentiment detecting task. For improving the performance of Chinese social media sentiment analysis, we created a Chinese Internet slang lexicon (examples shown in Table 1). We manually extracted 448 frequent Internet slang terms from the Internet New Words Ranking List, Baidu Baike⁵, Wikipedia⁶ and social media systems such as Baidu Tieba⁷ and Weibo⁸ between 2010 and 2018, and stored them as Chinese Internet Slang Lexicon. After analysis we observed that the entries fall under seven following categories:

- Numbers: such as 233 (“laughter/lol”: Chinese use 233 to express “can’t stop laughing” because 233 is an emotional sign in a Chinese BBS site⁹ and the sign is the NO.233 in the list of all emoticons); 213 (“a person who is very stupid”); 520/521 (“I love you”).
- Latin alphabet abbreviations: Chinese users commonly use a QWERTY keyboard with pinyin enabled. Upper case letters are quick to type and require no transformation. (Lower case letters spell words). Latin alphabet abbreviations (rather than Chinese characters) are also sometimes used to evade censorship. Such as SB (“dumb cunt”); YY (“fantasizing/sexual thoughts”); TT (“condom”).
- Chinese contractions: e.g. *ren jian bu chai* (“life is so hard that some lies are better not exposed”: This comes from the lyrics of a song entitled “*Shuo Huang*” (“Lies”), by Taiwanese singer Yoga Lin. This slang reflects that some people, especially young people in China, are disappointed by reality); *lei jue bu ai* (“too tired for romance”: this slang phrase is a literal abbreviation of the Chinese phrase “too tired to fall in love anymore”. It originated from an article on the Douban website, a Chinese social networking service website allowing registered users to record information and create content related to film, books, music, recent events and activities in Chinese cities. The article was posted by a 13-year-old boy who grumbled about his single status and expressed his weariness and frustration towards romantic love. The article went viral on the Chinese Internet, and the phrase was subsequently used as a sarcastic way to convey depression when encountering misfortunes or setbacks in life); *gao da shang* (“high-end, impressive, and high-class”: a popular

⁵ <https://baike.baidu.com>

⁶ <https://en.wikipedia.org>

⁷ <https://tieba.baidu.com>

⁸ <https://www.weibo.com>

⁹ <https://www.mop.com>

meme used to describe objects, people, behavior, or ideas that became popular in late 2013).

- Neologisms: *diao si* (“loser”): The word *diao si* is used to describe young males who were born into a poor family and are unable to improve their financial status. People usually use this phrase in an ironic and self-deprecating way); *ye shi zui le* (“nothing to say”): it is a way to gently express your frustration with someone or something that is completely unreasonable and unacceptable); *dan shen gou* (“single dog”): a term which single people in China use to poke fun at themselves for being single).
- Phrases with altered or extended meanings: *hao* or *tu hao* (“vulgar tycoon”): This word refers to irritating online game players who buy large amounts of game weapons in order to be gloried by others. Starting from late 2013, the meaning has changed and now is widely used to describe nouveau riche people in China who are wealthy but less cultured.); *bei tai* (“spare tire”): A girlfriend or boyfriend kept as a “backup”, “plan B”, just in case of breaking up with the current partner).
- Puns and wordplay: 河蟹 (“river crab”): pun on 和谐, another Chinese characters pronounced *he xie*, meaning “harmony”).
- Slang derived from foreign language: 工口 (The word *gong kou* comes from the Japanese katakana *ero*, which translated from English “erotic” into the abbreviation of the katakana 工口チツク, meaning “sensual”).

4 Lexicon of Chinese Social Media Emoticons

In the real-life (offline) dialogue between human beings, besides tone changes, we usually express emotions with body language. In social networks, this can partially be achieved by using emoticons [1].

There are many unknown factors in constantly changing moods of human beings, but communication with emoticons has become a global phenomenon. On the other hand, because of different ethnic and cultural differences, misunderstandings when using facial emoticons is not uncommon [16]. We also have noticed previously mentioned humorous emotion in Weibo microblog entries containing emoticons which are often difficult to interpret as positive or negative. It seems that some emoticons are used just for fun, self-mockery or jocosity which expresses an implicit humor characteristic in Chinese culture. Emoticons seem to play an important role in expressing this kind of emotion. There is a high possibility that this phenomenon can cause a significant difficulty in sentiment detecting task, therefore we decided to build a lexicon of emoticons before adding them to our system for classifying emotions in Weibo.

When we collected microblog data, we discovered that Weibo emoticons are transformed by API into Chinese characters, for example, 😊 will be convert into [微笑] (“smile”). This provided us with the possibility of building Chinese emoticon lexicon. Therefore, we selected the 109 Weibo emoticons (see Figure 2) which can be transformed into Chinese characters, and converted them into textual features to create Chinese emoticon lexicon. Several examples are shown in Table 2.

Table 2. Examples of Chinese Emoticon Lexicon.


























Emoticon	Textual Feature	Emotion/Implication
	[微笑]	“smile”
	[可爱]	“lovely”
	[太开心]	“too happy”
	[鼓掌]	“applause”
	[嘻嘻]	“hee hee”
	[哈哈]	“ha-ha”
	[笑cry]	“face with tears of joy”
	[挤眼]	“wink”
	[馋嘴]	“greedy”
	[黑线]	“speechless/awkward”
	[汗]	“sweat”
	[挖鼻]	“nosepick”
	[哼]	“snort”
	[怒]	“anger”
	[委屈]	“upset/fell wronged”
	[可怜]	“pathetic”
	[失望]	“disappointment”
	[悲伤]	“sad”
	[泪]	“weep”
	[害羞]	“shy”
	[污]	“filthy”
	[爱你]	“love face”
	[亲亲]	“kissy face”
	[色]	“leer”
	[舔屏]	“lick screen”
	[憧憬]	“longing”
	[二哈]	“dog leash”
	[摊手]	“smugshrug”



Fig. 2. 109 Weibo emoticons which can be transformed into Chinese characters.

5 Machine Learning approaches

Inspired by above mentioned works on Internet slang and emoticons, in order to test the influence of them, we utilized both lexicons with several machine learning approaches, k-Nearest Neighbors (k-NN), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machine (SVM) for detecting humorous expressions on social media. We did not tested deep learning approaches as the data size was not sufficient.

In the first step, we add the Chinese slang lexicon and Chinese emoticon lexicon to segmentation tool for matching new words and emoticons. Then we use the updated tool to segment the sentences of large data set. Second, we apply the segmentation results into the word embedding tool for training word vectors. Next, we apply the word embedding model which considered Internet slang and emoticons to train a machine learning model with training data. Finally, we input testing data into machine learning model, and we can obtain the sentiment probability of a Weibo post which considers the effect of emoticons and Internet slang.

6 Experiments

In order to verify the validity of our proposed method, we performed series of experiments described below.

6.1 Preprocessing

Initializing word vectors with those obtained from an unsupervised neural language model is a popular method to improve performance in the absence of a large supervised

training set. For our experiment we collected a large dataset (7.6 million posts) from Weibo API from May 2015 to July 2017 to be used for calculating word embeddings. First, we deleted the images, and videos treating them as noise. Second, we applied Chinese Internet slang lexicon and Chinese emoticon lexicon into the dictionary of Python Chinese word segmentation module Jieba¹⁰. Next, we used Jieba to segment the sentences of the microblogs, and applied the segmentation results into the word2vec model [11] for training word vectors. The vectors have dimensionality of 300 and were trained using the continuous skip-gram model.

Next, we collected 3,000 Weibo posts containing the emoticons. To use these posts as our training data, we asked three Chinese native speakers to annotate them into two categories: “humorous”, and “non-humorous”. After one annotator labelled polarities of all posts, two other native speakers confirmed correctness of his annotations. Whenever there was a disagreement, all decided the final polarity through discussion.

6.2 Applied Classifiers

Logistic Regression Logistic regression model is confirmed to be used in many tasks such as document classification [22]. In Logistic regression model, we generally correct overfitting with regularization. Regularization adds a penalty term on model to reduce the freedom of the model. Hence, the model will be less likely to fit the noise of the training data and will improve the generalization abilities of the model. We train the model with L2 penalty regularization called Ridge regression in our experiments.

Support Vector Machine Support vector machine [3] is a supervised learning model with associated learning algorithms that analyzes data used for classification. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition, it uses kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. In our experiments, we used the radial basis function kernel.

Naïve Bayes Naïve Bayes classifier is based on applying Bayes theorem with strong independence assumptions between the features. Naïve Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a baseline method for text categorization [15], the problem of judging documents as belonging to one category or the other with word frequencies as the features. With appropriate pre-processing, it is competitive in text classification task with more advanced methods including support vector machines. In our experiments, we set the parameter of alpha to 0.01.

k-Nearest Neighbors In pattern recognition, the k-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. In both cases, the input

¹⁰ <https://github.com/fxsjy/jieba>

consists of the k closest training examples in the feature space [20]. The output depends on whether k-NN is used for classification or regression. The number of neighbors is set to 5 in our experiments.

Random Forest Random forests is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [6]. Random decision forests correct for decision trees habit of overfitting to their training set.

Decision Tree Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision tree is commonly used in operations research, specifically in decision analysis to help identify a strategy most likely to reach a goal, but is also a popular tool in machine learning [17].

6.3 Performance Test

Using trained word2vec model, we passed word vectors of training data into the machine learning models to train the model. We collected and annotated 300 Weibo entries with emoticons as a test set, deleted images, and videos. Then we used the above mentioned methods to calculate scores of the precision, recall and F1-score. We compared the results of humorous detecting by machine learning only, machine learning considering Internet slang only, and machine learning approaches considering emoticons only. The results are shown in Table 3, Table 4 and Table 5, respectively. The Table 6 introduces results of the experiment where both Internet slang and emoticons were used, and Table 7 shows the results of F1-score with above methods.

Table 3. Comparison results of machine learning without emoticons and Internet slang lexicon.

	Precision	Recall	F1-score
DT	56.21%	55.56%	55.88%
RF	60.26%	54.97%	57.49%
k-NN	63.48%	66.08%	64.76%
NB	59.17%	83.04%	69.10%
LR	60.16%	86.55%	70.98%
SVM	57.00%	100.00%	72.61%

The results show that considering Internet slang and emoticons. Limited to small annotated data, the precision of the humor / non-humor classification was relatively low, but by considering Internet slang and emoticons, the F1-score of each classifier outperformed previous method by 1.39% (LR), 2.13% (SVM), 2.90% (NB), 0.69% (k-NN), 0.84% (RF) and 3.89% (DT). Our proposed approach has improved the performance showing that low-cost, small-scale data labeling is able to outperform widely

Table 4. Comparison results of machine learning with Internet slang lexicon only.

	Precision	Recall	F1-score
RF	59.76%	57.31%	58.51%
k-NN	60.10%	69.59%	64.50%
DT	66.46%	63.74%	65.07%
NB	59.92%	83.04%	69.61%
LR	60.16%	88.30%	71.56%
SVM	57.00%	100.0%	72.61%

Table 5. Comparison results of machine learning with Chinese emoticons lexicon only.

	Precision	Recall	F1-score
RF	61.59%	54.39%	57.76%
DT	63.37%	63.74%	63.56%
k-NN	60.70%	71.35%	65.59%
NB	59.92%	83.04%	69.61%
LR	60.08%	88.89%	71.70%
SVM	58.44%	100.00%	73.77%

Table 6. Comparison results of machine learning with both emoticons and slang.

	Precision	Recall	F1-score
RF	61.29%	54.75%	58.33%
DT	62.50%	57.26%	59.77%
k-NN	61.37%	70.11%	65.45%
NB	63.60%	82.96%	72.00%
LR	62.30%	86.31%	72.37%
SVM	59.67%	100.00%	74.74%

Table 7. Comparison results of F scores between feature sets.

	Baseline	Slang	Emoticons	Both
RF	57.49%	58.51%	57.76%	58.33%
DT	55.88%	65.07%	63.56%	59.77%
k-NN	64.76%	64.50%	65.59%	65.45%
NB	69.10%	69.61%	69.61%	72.00%
LR	70.98%	71.56%	71.70%	72.37%
SVM	72.61%	72.61%	73.77%	74.74%

used state-of-the-art when emoticon and slang information is added to the learning process.

7 Considerations

In our proposed approach, we paid more attention to the emoticons and Internet slang in microblogs and investigated how adding these features separately and together influences the previously proposed method for recognizing humorous posts which are problematic when it comes to semantic analysis. Figure 3) shown an example of a microblog which was correctly classified by our proposed method as “humorous” while the baseline recognized it incorrectly as non-humorous. This post contains word 一颗赛艇 (*yi ke sai ting* which is a homophone of English word “exciting”). The baseline does not know this expression and the parser divides it as 一颗/赛艇 (*yi ke / sai ting* which means “a rowing boat”). When this expression is accompanied by 😊 emoticon, they both improve the performance of classification and predict the implicit humorous meaning.

Post: 理想生活, 一颗赛艇 😊
Pinyin: <i>Li xiang sheng huo, yi ke sai ting</i> 😊
Segmentation: 理想/生活/, /一颗赛艇/[微笑]
Translation: Ideal life, exciting. 😊

Fig. 3. Example of correct classification of humorous post.

Error analysis showed that some posts were wrongly predicted due to proper nouns missing in the parser’s dictionary which brought clearly negative impact on the results. In Figure 4 we show an example of such misclassification into “non-humorous” category annotated as “humorous” by annotators. Name of a ticketing website *Da mai wang* was parsed incorrectly, and one shifted character caused mis-recognition of humorous word. Weibo microblogs contain numerous ideograms deliberately altered from their everyday meaning, what makes them difficult to parse and match. We think that adding new named entities into the parser’s dictionary may significantly improve the results in the future. We observed that when emotions are expressed online, emoticons might play a greater role than it is usually considered, therefore we will experiment with weight of the emoticons in the future.

Post: 大麦网真会玩, 2.14给我发个5.21号演唱会的票务信息 🙄

Pinyin: *Da mai wang zhen hui wan, 2.14 gei wo fa ge 5.21 hao yan chang hui de piao wu xin xi 🙄*

Segmentation: 大麦/网真会/玩/, /2.14/给我发/个/5.21/号/演唱会/的/票务/信息/[摊手]

Translation: Damaiwang really knows how to live it up, they sent me a concert ticket information of May 21st on February 14 🙄

Fig. 4. Example of wrong classification into “non-humorous” category.

8 Conclusions and Future Work

In this paper, we proposed adding Chinese Internet slang and emoticons for automatic classification of humorous posts on social media platform Weibo in order to separate them from clearly positive and negative ones. We collected 448 frequent Internet slang expressions and created a slang lexicon, then we converted the 109 Weibo emoticons into textual features creating Chinese emoticon lexicon. To test the influence of slang and emoticons on sentiment analysis task, we utilized both lexicons with several machine learning-based classifiers, namely k-Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, Naïve Bayes and Support Vector Machine for detecting humorous expressions on Chinese social media. Our experimental results show that the proposed additions can significantly improve the F1-score for detecting humorous expressions which are difficult to polarize into positive-negative categories.

For improving the performance of the proposed method, in near future we are going to increase the size of both slang and emoticon lexicons to improve further classification results. Furthermore, we plan to add image processing for classifying stickers which also seem to convey rich emotional information. Our ultimate goal is to investigate how much the newly introduced features are beneficial for sentiment analysis by feeding them to a deep learning model which should allow us to construct a high-quality sentiment recognizer for wider spectrum of sentiment in Chinese language.

9 Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 17K00295.

References

1. Aldunate, N., González-Ibáñez, R.: An integrated review of emoticons in computer-mediated communication. *Frontiers in psychology* **7**, 2061 (2017)
2. Chen, Z., Xu, R., Gui, L., Lu, Q.: Combining convolution neural network and word sentiment sequence features for chinese text sentiment analysis. *Journal of Chinese Information Processing* (2015)
3. Cortes, C., Vapnik, V.: Support-vector networks. *Machine learning* **20**(3), 273–297 (1995)
4. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017)
5. Guibon, G., Ochs, M., Bellot, P.: From emojis to sentiment analysis. In: *WACAI 2016* (2016)
6. Ho, T.K.: Random decision forests. In: *Document analysis and recognition, 1995., proceedings of the third international conference on*. vol. 1, pp. 278–282. IEEE (1995)
7. Kim, Y.: Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014)
8. Kulkarni, V., Wang, W.Y.: Tfw, damngina, juvie, and hotsie-totsie: On the linguistic and social aspects of internet slang. *arXiv preprint arXiv:1712.08291* (2017)
9. Li, D., Rzepka, R., Ptaszynski, M., Araki, K.: Emoticon-aware recurrent neural network model for chinese sentiment analysis. In: *The Ninth IEEE International Conference on Awareness Science and Technology (iCAST 2018)* (2018)
10. Li, R., Shi, S., Huang, H., Su, C., Wang, T.: A method of polarity computation of chinese sentiment words based on gaussian distribution. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. pp. 53–61. Springer (2014)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013)
12. Moschini, I.: The” face with tears of joy” emoji. a socio-semiotic and multimodal insight into a japan-america mash-up. *HERMES-Journal of Language and Communication in Business* (55), 11–25 (2016)
13. Novak, P.K., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. *PloS one* **10**(12), e0144296 (2015)
14. Peng, H., Cambria, E., Hussain, A.: A review of sentiment analysis research in chinese language. *Cognitive Computation* **9**(4), 423–435 (2017)
15. Rish, I., et al.: An empirical study of the naive bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. vol. 3, pp. 41–46. IBM New York (2001)
16. Rzepka, R., Okumura, N., Ptaszynski, M.: Worlds linking faces – meaning and possibilities of contemporary pictograms. *Journal of the Japanese Society for Artificial Intelligence* (2017)
17. Sharma, P., Kaur, M.: Classification in pattern recognition: A review. *International Journal of Advanced Research in Computer Science and Software Engineering* **3**(4) (2013)
18. Tan, S., Zhang, J.: An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications* **34**(4), 2622–2629 (2008)
19. Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., Bao, Z.: A depression detection model based on sentiment analysis in micro-blog social network. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. pp. 201–213. Springer (2013)

20. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* **10**(Feb), 207–244 (2009)
21. Wu, L., Morstatter, F., Liu, H.: Slangs4d: Building and using a sentiment dictionary of slang words for short-text sentiment classification. *arXiv preprint arXiv:1608.05129* (2016)
22. Yu, H.F., Huang, F.L., Lin, C.J.: Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning* **85**(1-2), 41–75 (2011)
23. Zagibalov, T., Carroll, J.: Automatic seed word selection for unsupervised sentiment classification of chinese text. In: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. pp. 1073–1080. Association for Computational Linguistics (2008)
24. Zhang, C., Zeng, D., Li, J., Wang, F.Y., Zuo, W.: Sentiment analysis of chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology* **60**(12), 2474–2487 (2009)
25. Zhao, P., Jia, J., An, Y., Liang, J., Xie, L., Luo, J.: Analyzing and predicting emoji usages in social media. In: *Companion of the The Web Conference 2018 on The Web Conference 2018*. pp. 327–334. International World Wide Web Conferences Steering Committee (2018)
26. Zhuo, S., Wu, X., Luo, X.: Chinese text sentiment analysis based on fuzzy semantic model. In: *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2014 IEEE 13th International Conference on*. pp. 535–540. IEEE (2014)