Emotion Corpus Construction on Microblog Text

Lei Huang, Shoushan Li(✉), and Guodong Zhou

Natural Language Processing Lab, School of Computer Science and Technology,
Soochow University, Suzhou, China
lei.huang2013@gmail.com, {lishoushan,gdzhou}@suda.edu.cn

Abstract. The construction of emotion corpus is a basic task in emotion analysis, which aims to annotate the emotions, such as *joy*, *sadness*, and *angry*, expressed in the text. Previous studies generally focused on the basic emotions in their emotion taxonomies. We find that, however, many emotions expressed in microblog text are difficult to be mapped to basic emotions, which poses a great challenge to emotion annotation. To address this problem, this paper proposes a novel emotion taxonomy which contains both basic emotions and complex emotions for annotating emotions expressed in microblog text. Specially, the basic emotions include four emotion classes, i.e., *joy*, *angry*, *sadness*, and *fear* while the complex emotions include *positive* emotions, *neutral* emotions, and *negative* emotions. Experimental results demonstrate that the emotion annotation with the proposed emotion taxonomy achieves a much higher consistency, providing a good foundation for further emotion recognition task.

Keywords: Emotion corpus · Emotion analysis · Consistency

1 Introduction

With the emergence and development of Web2.0, the amount of information on the web is increasing rapidly. Microblog, or Weibo in Chinese, has become one of the most popular internet applications due to its expression features like more free and convenient among social media. Under this background, text emotion analysis provides important technical means for automatic processing and analyzing mass data in the microblogs. The text emotion analysis has become a hot research task in computational linguistics due to its wide potentials applications, e.g. public opinion monitor, intelligent advertisement, emergency warning, etc. In addition, emotion analysis can also help other researches fields, such as psychology, sociology, finance.

Emotion refers to the inner psychological reactions and feelings, such as *joy*, *angry*, *sadness*, *fear*, etc. Emotion analysis aims to automatic identify author's mood, psychological reaction and emotional state expressed by a piece of text. Generally, emotion analysis has two basic tasks: emotion recognition and emotion classification (Aman and Szpakowiczm, 2007). The existing approaches of emotion analysis are commonly based on corpus classification methods. Corpus construction is the basic technological means in natural language processing. The key of corpus construction is to build practical annotation system. In the construction of emotion analysis corpus, the emotion taxonomy is foundation and difficulty of emotion analysis research.

Q. Lu and H.H. Gao (Eds.): CLSW 2015, LNAI 9332, pp. 204–212, 2015.
DOI: 10.1007/978-3-319-27194-1_21

Although there are large divergences in categories and quantities of the emotion, researchers from different research fields all agree that emotions can be divided into two categories, i.e., basic emotions and complex emotions (Arnold，1960；Ortony and Turner, 1990). In previous researches, the basic emotions classes are divided into 4, 6, 8, 10 or even more than 20, which is mainly because of the complexity and variability of emotions, and there is no complete and systematic understanding in the academic world. Such as Gray (1982) considered the basic emotions are *joy*, *rage*, *anxiety*, *fear*; Ekman (1982) divided the basic emotions into *joy*, *sadness*, *anger*, *fear*, *disgust*, *surprise*; Pultchik (1980) proposed 8 basic emotions classes, i.e., *acceptance*, *anger*, *anticipation*, *disgust*, *joy*, *fear*, *sadness*, *surprise*.

The complex emotions are difficult to list them all, common ones include *gratitude*, *pride*, *guilt*, etc. The existing text emotion taxonomies tend to map a variety of complex emotions to the basic emotions, such as we can map annoyance to *sadness*, satisfactory to *joy*. However, in the practical annotation process, all the complex emotions were mapped to the basic emotions is not easy task. For example, see the following two sentences:

(a) Chinese: *感谢你们这一路的陪伴。*
Pinyin: *gǎn xiè nǐ men zhè yí lù de péi bàn.*
English: *Thank you for you company.*
(b) Chinese: *今天父亲节，祝爸爸身体健康！*
Pinyin: *jīn tiān fù qīn jié, zhù bà ba shēn tǐ jiàn kāng.*
English: *Today is father's day, I wish my father good health!*

From the above two examples, we can see that the sentence (a) expresses gratitude emotion and the sentence (b) reveals blessing emotion. These two emotions are complex emotion and are difficult to be mapped to the basic emotions. In addition, obviously, the two sentences were labeled as *emotionlessness* is unreasonable. Therefore, it is very difficult to only use some basic emotions to map all emotions for annotating corpus.

This paper aims to construct emotion corpus on microblog text. We propose a novel emotion taxonomy which contains basic emotions and complex emotions for annotating emotions expressed in microblog text. The basic emotions include four emotion classes, i.e., *joy*, *angry*, *sadness*, and *fear*. These emotions are basically all relevant researches consistent with the basic emotions (Ortony and Turner, 1990). Emotions except for the basic emotions are complex emotions. The complex emotions include *positive* emotions, *neutral* emotions, and *negative* emotions. In addition, we have set *emotionlessness* category. Empirical studies on emotion corpus demonstrate that our proposed emotion taxonomy improves the efficiency of annotating corpus and achieves a much higher consistency.

The remainder of this paper is organized as follows. Section 2 overviews the related work on emotion analysis and emotion corpus construction. Section 3 explores the description and construction methods of emotion corpus. Section 4 presents the statistics and analysis of constructed corpus. Finally, Section 5 gives the conclusion and future work.

2 Related Work

Over the last decade, emotion analysis has been a hot research topic and involve various aspects, such as emotion resource creation (Wiebe et al., 2005; Quan and Ren, 2009; Xu et al., 2010), writer's emotion vs. reader's emotion analysis (Lin et al., 2008; Liu et al., 2013), emotion cause event analysis (Chen et al., 2010), document-level emotion classification (Alm et al., 2005; Li et al., 2014) and sentence-level or short text-level emotion classification (Tokushisa et al., 2008; Bhowmick et al., 2009; Xu et al., 2012). This work focuses on construction of emotion analysis corpus.

Mishne (2005) selected 132 common moods and constructed a corpus of 815494 blog posts in Livejournal. Pak and Paroubek (2010) collected a corpus of text posts and formed a dataset of three classes: positive sentiments, negative sentiment, and a set of objective texts by using Twitter API. Quan and Ren (2009) proposed an emotional expression space model and described a relatively fine-grained annotation scheme and annotated emotion in text. In aspects of Chinese emotion corpus, Xu et al. (2008) adopted emotion classification contains 7 coarse classes, 22 fine classes for emotion annotating on sentence level in primary school textbooks, screenplays, and completed 39488 sentences of annotation. Yao et al. (2014) used seven emotion classes for annotating emotions on microblog level and sentence level on Sina microblog text, this corpus consists of 14000 microblogs, totaling 45431 sentences.

3 Emotion Corpus Construction

3.1 Datasets and Pre-processing

Up to now, the researches on emotion corpus construction on Chinese microblog are relatively small, and lack of related public corpus. The primary task of this paper is to collect relevant data. We download microblogs from the website http://t.qq.com. The data were collected by using the Tencent Microblog API, and contain 260000 users' personal information and their tweets. Specially, microblogs of each user were collected at most 210. The types of microblog are divided into seven types, i.e., original published, retweet, private message, reply, empty reply, comment, mention, according to the Type values of Tencent Microblog API. In this emotion taxonomy, we only focus on the emotion status of microblog publisher. Thus, we only choose the original published tweet to construct emotion corpus. Currently, this emotion annotated corpus consists of 150 users, totaling 15540 microblogs.

3.2 Emotion Taxonomy

The construction of annotation system of emotion analysis corpus is the mainly emotion classification problem. Because of the complexity and variability of emotion, and the number of emotion make it difficult for researchers to determine its type. The existing emotion classification systems are inconsistent. The granularity of emotion taxonomy is too detailed, not only increases the difficulty of annotation work, but also reduces the consistency of annotation. If the granularity of emotion taxonomy is too coarse, it is difficult to cover some common emotions. Therefore, this paper proposes a novel emotion taxonomy which both basic emotions and complex emotions. Specially, the basic emotions include four emotion classes, i.e., *joy*, *angry*, *sadness*, *fear*. While complex emotions contain three emotion classes, i.e., *positive* emotions, *neutral* emotions, and *negative* emotions. In addition, we have set *emotionlessness* category. Figure 1 illustrates the emotion taxonomy in detail as follows.

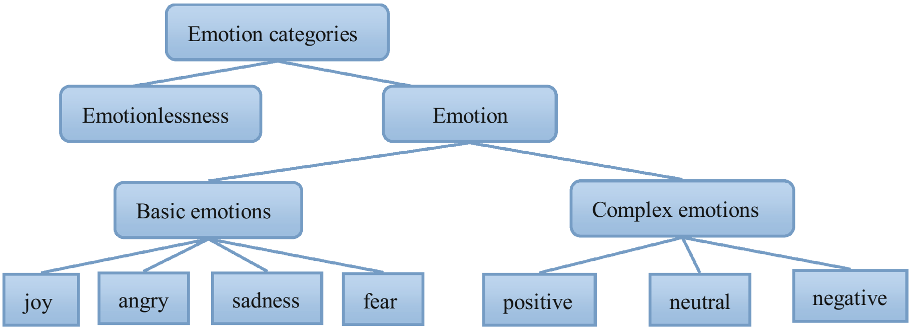

Fig. 1. Emotion taxonomy

3.3 Annotation Guidelines

The annotation guidelines can help to reduce error operation, improve the speed of annotation and increase the consistency of annotation. A microblog can have a maximum of 140 characters, colloquial and brief expression, have more network language. Moreover, Microblog platform also provides rich emoticons, such as “🥺”, the text elements corresponding to “∨*大哭*”[*dà kū*](*crying*). These emoticons often represent user’s emotional tendency. Therefore, the annotation guidelines were designed by combining with the characteristics of the microblog for constructing emotion corpus on microblog text is of great significance. The guidelines of this paper are as follows:

1) Annotating process: In the annotation process, we should follow the basic emotions before complex emotions, namely, for a microblog text containing emotion, we firstly judge whether the emotion belongs to the basic emotions, if it belongs to basic emotions, we choose the basic emotions for annotation. Otherwise, judge the type of complex emotions which it belongs to, we choose the appropriate complex emotions for annotating. For example,
   Chinese: *最近好烦啊*
   Pinyin: *zuì jìn hǎo fán ā*
   English: *It’s been a very annoying.*

   Obviously, the text contains emotion. We firstly should judge whether the emotion belongs to the basic emotions, after finding no appropriate type of the basic emotions, therefore, choose *negative* emotion from the complex emotions category for annotating.

2) The number of Annotators: Each microblog was respectively annotated by two annotators, if the results of annotation are consistent, this results as the final results. If the results are inconsistent, so that the third annotator take part in annotation work, the consistent results as the final results.
3) Emoticon microblog: If a microblog text only contains emoticon, then directly annotate according to the meaning of emoticon. For instance, “\/大哭”[dà kū](*crying*) annotate *sadness*. If the meaning of emoticon is different from the meaning of the microblog text semantics, we should label microblog according to microblog text semantics. For example,

   Chinese: *火大，去了邮局几趟都没有把我的支付宝弄好，一群笨蛋，* 🙂

   Pinyin: *huǒ dà qù le yóu jú jǐ tang dōu méi yǒu bǎ wǒ de zhī fù bǎo nòng hǎ o, yì qún bèn dàn,* 🙂

   English: *I was very angry, went to the post office several times, the post office employees didn't make my Paypal ready, a group of idiots.* 🙂

   In this case, the microblog expresses *anger* emotion, but uses a smile emoticon, we should label as *anger* according to microblog semantics.
4) Multiple label: If a microblog text contains multiple emotions, we should respectively label each emotion. For example,

   Chinese: *爸爸，我们之间不仅是父子，还是师生，当你的学生真好，祝你快乐一生！*

   Pinyin: *bà ba wǒ men zhī jiān bù jǐn shì fù zǐ, hái shì shī sheng, dāng nǐ de xué sheng zhēn hǎo, zhù nǐ kuài lè yì sheng.*

   English: *Dad, we are not only father and son, or teachers and students, it's good to be as your student, I wish you a happy life.*

   This text contains two emotion types, happy and bless, so we label as *joy* and *neutral*.
5) Joke microblog: if the content of microblog text is a joke or the author joking, we uniformly label as *joy*. For example,

   Chinese: *我买了一只仓鼠和一个笼子，有天和朋友抱怨：你说，这笼子竟然比仓鼠还贵。朋友说：难道你认为你会比现在的房价高吗？*

   Pinyin: *wǒ mǎi le yī zhī cāng shǔ hé yī gè long zi, yǒu tiān hé péng yǒu bào yuan: nǐ shuō zhè long zi jìng rán bǐ cāng shǔ hái guì. Péng yǒu shuō nán dào nǐ rèn wéi nǐ huì bǐ xiàn zài de fang jià gāo ma ?*

   English: *I bought a hamster and a cage. One day with friends complained: you said, this cage is more expensive than the hamster. Friend said: do you think you will be higher than the current house price?*
6) Swearing microblog: If the content of microblog text contains dirty words, we can give priority to determine whether it contains *anger* emotion. For example,

   Chinese: *草泥马的铁路公司，尼玛的车就不能准时到站一回。*

   Pinyin: *cǎo ní mǎ de tiě lù gong sī, ní mǎ de chē jiù bù néng zhǔn shí dào zhàn yī huí.*

   English: *Fuck the railroad, fuck the train can't arrive on time one time.*
7) Share microblog: If the content of a microblog text is about the shares of games, songs, shopping ads, etc., we should label as *emotionlessness*.

Chinese: *龙是桀骜的力量。我参加了《剑灵》人龙灵天种族投票，你的种族选好了吗？还有免费参加#剑灵不删档激活码抽奖#*
Pinyin: *long shì jié ào de lì liang, wǒ cān jiā le jiàn líng rén long líng tiān zhǒng zú tóu piào, n ǐ de zhǒng zú xuǎn hǎo le ma? Hái yǒu miǎn fèi cān jiā jiàn líng bù shān dàng jī huó m ǎ chōu jiǎng*
English: *The dragon is an unruly power. I participated in four races (Gon, Jin, Yun, Lyn) vote of BNS (Blade and Soul). Did your race choose? There are free to participate in activity for obtaining activation code.*

4 Analysis of Corpus

As discussed in Section 3.3, we annotate the emotional microblog texts. The annotated corpus consist of 150 users published 15540 microblogs. Table 1 summarizes the distribution of emotion on the corpus. Table 2 shows the distribution of emotion categories expressed in emotional microblog text.

Table 1. The distribution of emotion

	Emotional microblog		Emotionless microblog
	Single label	Multiple label	
amount	6431	117	8992
proportion/%	41.38%	0.76%	57.86%
total/%	42.14%		57.86%

Table 2. The distribution of emotional texts in each emotion category

Categories	Amount	Proportion/%
joy	1038	15.56%
angry	472	7.08%
sadness	581	8.71%
fear	94	1.41%
Positive complex emotion	1178	17.66%
Neutral complex emotion	1131	16.96%
Negative complex emotion	2175	32.61%
ALL	6669	100%

From the Table 1, we can see that the proportion of emotional and emotionless microblog is about 1:1.36. It is worth nothing that the proportion of multiple label microblog text is small compared with single microblog text. This is mainly due to our proposed emotion taxonomy in this paper can sufficient reflect the distinction between categories, and has a good degree of discrimination.

Obviously, the distribution is a bit imbalanced in Table 2. In basic emotions, the proportion of *joy* is the biggest, and the proportion of *anger* is the smallest. In complex emotions, the proportion of negative is the most.

This paper adopts kappa value to measure the consistency of annotating. From the two aspects of the consistency measurement, one is whether microblog text contains emotion, and the other is the emotion category. The consistencies of our annotation on the microblog text are given in Table 3.

Table 3. Analysis of consistency

	Kappa value
Emotion or emotionlessness	0.7186
Emotion category	0.7156

From the results of Table 3, we can see that the consistency of emotion category reach 0.7156, while the consistency of emotion or emotionlessness is about 0.7186. This results demonstrate our annotation work achieve a much higher consistency.

In order to reflect the advantages of our proposed emotion taxonomy, we randomly select 492 microblogs that 6 users published and use the basic emotions to annotate. Based on this, we carry out some statistics and analyses.

Table 4 shows the number of inconsistencies under the two emotion taxonomy. One is only containing four basic emotions, i.e., *joy*, *angry*, *sadness*, *fear* while another is our proposed emotion taxonomy which contains both basic emotions and complex emotions. The number of inconsistency is the number of inconsistent microblog text that different annotators label under the same samples.

Table 4. The statistic of the two emotion taxonomies

User	Only four basic emotions	Our proposed emotion taxonomy
1	31（26.27%）	3 （2.54%）
2	67（46.53%）	27（18.75%）
3	9 （30.00%）	4 （13.33%）
4	22（25.29%）	16（18.39%）
5	34（44.74%）	17（22.37%）
6	20（54.05%）	4 （10.81%）

From the statistics of Table 4, we can find that in our proposed emotion taxonomy, the proportion of inconsistency of microblog texts, the worst is 22.37%, and the best is 2.54%. While in the only four basic emotions, under the same sample, the proportion of microblog text annotation, the best is 25.29%. In contrast, this paper proposed emotion taxonomy can better distinguish different emotion categories, therefore, achieves a much higher annotation results.

5 Conclusion and Future Work

In this paper, we propose a novel emotion taxonomy which contains both basic emotions and complex emotions in order to construct emotion corpus on microblog text. Up to now, we have annotated 15540 microblogs that 150 users published. On this basis, we have carried on statistic and analysis for results of annotation. The results show that the emotion taxonomy that we proposed can well construct emotion corpus on microblog text, and the consistency of the annotation is about 0.72. In addition, we find that compared with the traditional emotion classification system, the emotion taxonomy can improve the quality and efficiency of annotating corpus.

In the future work, we would like to annotate corpus work in order to further expand the scale of corpus, and apply our annotated emotion corpus in more applications, e.g. microblog emotion classification.

Acknowledgments. This research work has been partially supported by two NSFC grants, No.61375073 and No.61273320, one the State Key Program of National Natural Science of China No.61331011.

References

1. Alm, C., Roth, D., Sproat, R., Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of EMNLP, pp. 579–586 (2005)
2. Aman, S., Szpakowicz, S.: Identifying expressions of emotion in text. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 196–205. Springer, Heidelberg (2007)
3. Arnold, M.: Emotion and Personality. Columbia University Press, New York (1960)
4. Bhowmick, P.K., Basu, A., Mitra, P., Prasad, A.: Multi-label text classification approach for sentence level news emotion analysis. In: Chaudhury, S., Mitra, S., Murthy, C.A., Sastry, P.S., Pal, S.K. (eds.) PReMI 2009. LNCS, vol. 5909, pp. 261–266. Springer, Heidelberg (2009)
5. Chen, Y., Lee, S., Li, S., Huang, C.: Emotion cause detection with linguistic constructions. In: Proceedings of COLING, pp. 179–187 (2010)
6. Ekman, P., Friesen, V., Ellsworth, P.: What emotion categories or dimensions can observers judge from facial behavior? In: Ekman, P. (ed.) Emotion in Human Face, pp. 39–55. Cambridge University Press, New York
7. Gray, A.: The neuropsychology of anxiety. Oxford University Press, Oxford (1982)
8. Li, C., Wu, H., Jin, Q.: Emotion classification of chinese microblog text via fusion of bow and evector feature representations. In: Zong, C., Nie, J.-Y., Zhao, D., Feng, Y. (eds.) NLPCC 2014. CCIS, vol. 496, pp. 217–228. Springer, Heidelberg (2014)
9. Lin, K., Yang, C., Chen, H.: Emotion classification of online news articles from the reader's perspective. In: Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, pp. 220–226 (2008)
10. Liu, H., Li, S., Zhou, G., Huang, C., Li, P.: Joint modeling of news reader's and comment writer's emotions. In: Proceedings of ACL, pp. 511–515 (2013)
11. Mishne, G.: Experiments with mood classification in blog posts. In: Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access (2005)
12. Ortony, A., Turer, T.: What's Basic about Basic Emotions? Psychological Review, 315–331 (1990)
13. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of Language Resource and Evalutation Conference, pp. 1320–1326 (2010)
14. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Plutchik, R., Kellerman, H. (eds.) Emotions: Theory, Research, and Experience. Theories of emotion, vol. 1, pp. 3–31. Academic Press, New York
15. Quan, C, Ren, F.: Construction of a blog emotion corpus for Chinese emotional expression analysis. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1446–1454 (2009)

16. Tokuhisa, R., Inui, K., Matsumoto, Y.: Emotion classification using massive examples extracted from the web. In: Proceedings of COLING, pp. 881–888 (2008)
17. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation **39**, 65–210 (2005)
18. Xu, G., Meng, X., Wang, H.: Build Chinese emotion lexicons using a graph-based algorithm and multiple resources. In: Proceedings of COLING, pp. 1209–1217 (2010)
19. Xu, J., Xu, R., Lu, Q., Wang, X.: Coarse-to-fine sentence-level emotion classification based on the intra-sentence features and sentential context. In: Proceedings of CIKM-2012, poster, pp. 2455–2458 (2012)
20. Xu, L., Lin, H., Zhao, J.: Construction and Analysis of Emotional Corpus. Journal of Chinese Information Processing **22**(1), 116–122 (2008)
21. Yao, Y., Wang, S., Xu, R., et al.: The Construction of an Emotion Annotated Corpus on Micro blog Text. Journal of Chinese Information Processing **28**(5), 83–91 (2014)