Emotion Corpus Construction Based on Selection from Hashtags

Minglei Li, Yunfei Long, Qin Lu, Wenjie Li
Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
E-mail: {csmli, csylong, csluqin, cswjli}@comp.polyu.edu.hk

Abstract

The availability of labelled corpus is of great importance for supervised learning in emotion classification tasks. Because it is time-consuming to manually label text, hashtags have been used as naturally annotated labels to obtain a large amount of labelled training data from microblog. However, natural hashtags contain too much noise for it to be used directly in learning algorithms. In this paper, we design a three-stage semi-automatic method to construct an emotion corpus from microblogs. Firstly, a lexicon based voting approach is used to verify the hashtag automatically. Secondly, a SVM based classifier is used to select the data whose natural labels are consistent with the predicted labels. Finally, the remaining data will be manually examined to filter out the noisy data. Out of about 48K filtered Chinese microblogs, 39k microblogs are selected to form the final corpus with the Kappa value reaching over 0.92 for the automatic parts and over 0.81 for the manual part. The proportion of automatic selection reaches 54.1%. Thus, the method can reduce about 44.5% of manual workload for acquiring quality data. Experiment on a classifier trained on this corpus shows that it achieves comparable results compared to the manually annotated NLP&CC2013 corpus.

Keywords: hashtag; emotion corpus construction; high-quality label selection

1. Introduction and Related Works

Emotion classification for text from the social media (such as Twitter, Sina Weibo) is becoming more and more important because emotions expressed in these media can have a social impact in the society. Many supervised learning methods have been employed to solve this problem. However, supervised methods require a large amount of labelled training data. Obtaining labelled data, if annotated manually, is very time-consuming especially for emotion analysis where training data can be quite skewed for multiple classes.

There are a number of emotion corpora from previous research works. For the English language, SemEval2007 (Strapparava and Mihalcea, 2007) consists of only 1,250 news headlines labelled with the six Ekman emotion labels. The ISEAR dataset (Scherer and Wallbott, 1994) consists of 7,666 sentences generated by participants through questionnaires. The Affect Dataset (Alm 2009) consists of more than 15,000 sentences from fairy tales with five emotion labels. For Chinese, the Ren-CECps (Quan and Ren, 2010) emotion corpus consists of 1,487 documents and 35,096 sentences from web blogs with eight emotions. The limitations of the above datasets are that they are either too small in size or not proper for social media text analysis. Another social media orientated Chinese corpus NLP&&CC 2013 (Yuanlin et al., 2014) consists of 14,000 microblogs and 45,431 sentences from microblogs with 8 labels (including "none" label, meaning no emotion) through manual annotation only 7,300 microblogs contain emotions and the size is still quite small. Those corpora are not suitable for large-scale emotion analysis.

Many research studies employ distant supervision and they take advantage of large amounts of text available in social media to investigate automatic methods to obtain labelled data (Wang et al., 2012; Tang et al., 2013; Mohammad and Kiritchenko, 2014). In these works, naturally annotated text features such as hashtags (the term inserted between two characters "#" by the author, called "topic" in Sina Weibo), emoticons and emoji characters in microblogs are automatically extracted from data and directly served as labels after some simple rule-based selection. It is possible to build a large-sized corpus using this method. The main issue with this method is that naturally annotated text naturally contains noise. Without appropriate methods to filter out noisy data would make the data less useful as it affects the performance in its usage.

Take the following text as an example, “在你闲的时候，玩玩转发微博，未必不是一种乐趣！！！#无聊# (When you are not busy, playing with microblog retweet may be fun! #boring#)”. From the text we can infer that the emotion is “happy”. But, the author uses a negative hashtag “boring". This example indicates that the text content is inconsistent with the naturally labelled hashtag. If this data is used as training data, it will obviously mislead machine learning algorithms as classifiers.

In this study, we present our work on how to make use of the naturally labelled data effectively and at the same time try to eliminate noisy data to reduce the detriment of the noise. The training data we try to obtain is from social media for the purpose of building an emotion corpus for emotion analysis. The basic idea is to use a multiple stage method to first select high-quality naturally labelled data automatically and then, use experts to manually examine data in the remaining set for correct annotation. Commonly used natural labels include emoticon, emoji and hashtags. One advantage of hashtag over emoticon and emoji is that we can search the microblogs based on a given hashtag. So we adopt hashtag as the natural label in this study. Results show that out of about 48K filtered Chinese microblogs (from about 173K raw microblogs), 39k microblogs are selected to form the final corpus with the Kappa value reaching over 0.92 for the automatically selected part and over 0.81 for the manual part. The proportion of the automatically selected part is 54.1% and the manual part is 45.9%. Thus, the method can reduce about 44.5% workload compared to the manual workload for acquiring high-quality data. Experiment on a classifier using this corpus as training data shows that it achieves comparable results compared to the classifier trained on the manually annotated NLP&CC2013 corpus.

The rest of the paper is organized as follows. Section 2 presents the methodology of constructing an emotion corpus based on naturally annotated hashtags. Section 3 analyzes the obtained emotion corpus. Section 4 gives conclusion and future work.

2. Emotion Corpus Construction

In this study, we focus on Chinese Microblog emotion corpus construction through hashtags, which are called “topics” in Sina Weibo. The original data are crawled from Sina Weibo through Sina Weibo Topic API[1]. For the emotion model, we keep it consistent with the NLP&CC2013 corpus which adopts seven emotion labels: *like, disgust, happiness, sadness, anger, surprise, fear*. The whole construction procedure is shown in **Figure 1**. First we select emotional seed hashtag words and based on these seed words to crawl microblog with hashtags. Then a simple rule based pre-processing is performed on the raw data. The pre-processed data then goes through a lexicon based selection (Part1) and a SVM classifier based selection (Part2). The remaining data is then manually examined in Step 5 to assure the label quality (Part3). Previous works on natural data selection only focus on steps 0, 1 and 2(marked by the blue box) to construct an emotion corpus. Our work goes further to include step 3, 4 and 5(marked in the red box). Each step will be discussed in detail in the following sections.

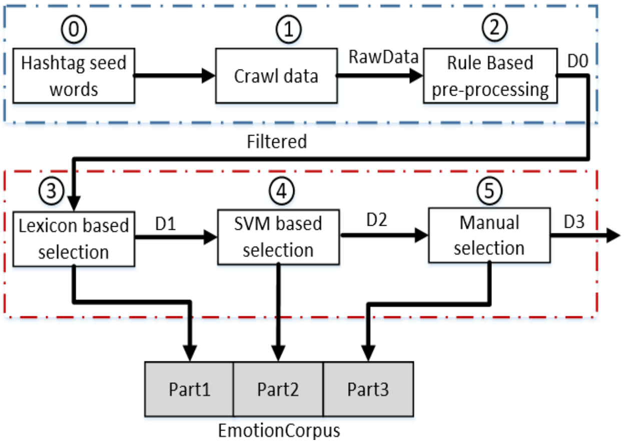

Figure 1 Emotion Corpus Construction Framework

2.1 Seed Hashtag Selection

The first step is to crawl data from Sina blogs. We manually select a set of seed topics for different emotion labels, as listed in **Table 1**. These categories follow the work of Xu (Xu et al., 2008) in which words are manually classified into different emotion categories. Based on these seed words, we crawl 173,958 microblogs with hashtags (denoted as **RawData**).

Emotion Label	Seed word number and examples
Like	9: “给力 (helpful)” “可爱 (lovely)” “奋斗 (strive)” “喜欢 (like)” “赞 (appraise)” “爱你 (love you)” “相信 (believe)” “鼓掌 (applaud)” “祝愿 (hope)”
Disgust	9: “无聊 (boring)” “烦躁 (agitated)” “嫉妒 (jealous)” “尴尬 (embarrassment)” “讨厌 (dislike)” “恶心(disgusting)” “怀疑 (suspect)” “烦闷 (bored)” “厌恶 (disgust)”
Happiness	20: “快乐(happy)” “幸福 (happy)” “哈哈 (ha-ha)” “爽 (so high)” “感动 (moved)” “开心 (joy)” “嘻嘻 (happy)” “高兴 (happy)” “亲亲 (kiss)” “欢喜 (happy)” etc.
Sadness	27: “伤不起 (can’t bear the hurt)” “郁闷 (sadness)” “哭 (cry)” “失望 (disappointed)” “心塞 (heart hurt)” “难过 (sadness)” “思念 (long for)” etc.
Anger	27: “妈的 (fuck)” “无语 (speechless)” “气愤 (angry)” “恼火 (anger)” “tmd” “你妹的 (your sister)” etc.
Surprise	16: “神奇 (miracle)” “惊呆了 (shocked)” “不可思议 (inconceivable)” “天哪 (my god)” “大吃一惊 (shocked)”
Fear	11: “害怕 (fearful)” “紧张 (nervous)” “心慌 (nervous)” “害羞 (shy)” “囧 (embarrassed)” etc.

Table 1 Hashtag Seed Words

2.2 Data Preprocessing

Since microblog is one kind of social media texts, the crawled data contain noise. In Step 2, pre-processing is conducted on the **RawData** based on the following manual rules.

1. Remove text that contains less than 3 words excluding the hashtag and URL.
2. Remove duplicated microblogs in a discussion chain.
3. Remove microblogs that contain URL.
4. Remove text whose hashtag is in the middle of the microblog.
5. Remove forwarded microblogs.
6. Remove text that is not in Chinese.
7. Remove text that contains more than one hashtag.
8. Remove text that contains quotes because such text is more likely to be dialogs, such as “讲个 故事：“从前 有个 太监 ………” 有人 耐不住 问：“下面呢？” 继续 讲故事：“下面？没了 啊……” (I told you a story, there is a eunuch...... Someone cannot hold their patient and ask, what's the second part of this story, I counter, second part? There is no second part to him. (Has been castrated.))”, which is just a joke in Chinese. There are many microblogs like this.
9. Convert traditional Chinese to Simplified Chinese.
10. Remove abnormal hashtags, such as “神奇 (amazing)”, which is the name of a movie. This is performed through manual review of the raw data.
11. Remove microblogs that contain more than one hashtags.
12. Remove microblogs whose hashtag is not at the start or the end of the text. This is because some hashtags are used as a part of the text content and cannot reflect the whole emotion of the text. For example, in the

[1] http://open.weibo.com/wiki/2/search/topics

sentence “我好#伤心#啊 (I am so #sad#)”, “sad” is the content of the text, if treat it as a hashtag and remove it from the text, the text is incomplete.

Through the above preprocessing, 48,305 (27.77%) microblogs are kept out of 173,958, indicating that 72.23% microblogs are noisy data. The hashtags in the cleaned microblogs (noted as the **D0** dataset) are converted to the corresponding emotion labels (denoted as ***natural labels***) based on **Table 1** and the text is segmented by the Chinese segmentation tool Jieba[2] for further processing.

2.3 Emotion Lexicon Based Selection

Since the natural labels may not be accurate as stated in introduction part, we use verification method in Step 3. We select the natural label based on an emotion lexicon counting strategy. The algorithm is shown below in **Algorithm 1**. Given one segmented text, we count words that actually in the emotion lexicon and use the emotion with a maximum count as the verified label of this text. If several emotion counts are equal, these emotion labels are all regarded as verified labels. If the original natural label is in the verified label set, we regard it as a high-quality label and add it to the selected high-quality dataset *H*. Otherwise, they will be included in the set for further processing. The emotion lexicon used is DUTIR[3] plus a collected popular internet words. After the lexicon based selection, 14,197 microblogs with the high-quality label are obtained (denoted as **Part1**) and 34,108 are left (denoted as **D1**) for further processing.

Algorithm 1: The Lexicon Based Algorithm

Inputs:
W = $[w_1, w_2, \ldots, w_m]$: Segmented text with *m* words.
y_o: the natural label of text W from the hashtag.
S = $[s_1, s_2, \ldots, s_n]$: Emotion lexicon for *n* emotions.
Y = $\{y_1, y_2, \ldots, y_n\}$:The emotion label set

Output:
H: The selected high-quality dataset.
N: The left dataset for further processing.

Procedure:

```
Set C = $[c_1, c_2, \ldots, c_n]$ where $c_i = 0$,
for w in W:
    for $s_i$ in S:
        if w in $s_i$:
            $c_i = c_i + 1$
max_c = argmax(C)
for $c_i$ in C:
    if $c_i$ == max_c:
        add $y_i$ $to$ $y$
if $y\_o$ in y:
    add W to $H$
else: add W to $N$
```

The lexicon based selection helps to identify text that contains explicit emotional words or emotion affinity words. Text that express emotion through word combination cannot be classified by a lexicon, such as “今天我这里又没有水了～～～(Today there is no water again in my place)” which expresses sadness through the combination of “no” and “water”. In such situation, we employ machine learning based selection in Step4.

2.4 SVM Based Selection

Step 4 uses a Support Vector Machine(SVM) (Suykens and Vandewalle, 1999) based selection methods. The basic idea is that a classifier is trained first based on the available high-quality emotion corpus and then, we use this classifier to predict the remaining data from Step 3. If the predicted label is the same as the original natural label, we regard it as a high-quality label and put it in **Part2** data. Otherwise, it is put to **D2** set for further processing. The features used in SVM is bag-of-words with stop words removed. SVM is implemented in Liblinear (Fan et al. 2008) as the classifier, which is widely used in text classification. The training data is from NLP&CC2013[4]. After selected by SVM, 7,228 microblogs are obtained as **Part2** data, and 26,820 are in **D2**.

2.5 Manual Annotation Based Selection

Since the remaining D2 data cannot be classified by the previous automatic steps, we ask one trained annotator to manually annotate it. The annotation rules are as follows:

1. Only consider the emotion of the author.
2. If the author describes something with positive or negative words, we claim that the author expresses “like” or “disgust” emotion.
3. Each microblog may contain several sentences, we only consider the emotion of the whole text. For example, the microblog “今天出门上班摔了一跤，不过还好碰到了个大帅哥把我带到了公司 (Today I fell down when I went to work. Fortunately, a handsome guy brought me to my office.)”, which expresses “sadness” in the first part and “happiness” in the second part. In this case, we set “happiness” as the major emotion label.
4. Each microblog can be labeled with at most two emotions. If one of them matches the natural label, we deem the natural label as a high-quality label.
5. For text that is meaningless which is not discarded in preprocessing, we discard it, such as “是良好的健康加上糟糕的记性. (…good health plus bad memory)” where the original text is “#幸福#是良好的健康加上糟糕的记性. (#Happiness# is good health plus bad memory)”. As we can see, even though the hashtag “happiness” is at the start of the microblog, it is part of the content, which cannot be simply recognized by the pre-processing.
6. For text that is judged no emotion, we label it as “**none**”.

Since only one annotator is trained to perform the annotation, there is a chance that the manual label is incorrect, and there is also a chance that the natural label is incorrect. But the chance of both labels are incorrect is much lower. So we only reserve those whose natural labels are consistent with the manual label and discard the rest. Finally, 18,236 microblogs are obtained in **Part3** dataset. The remaining 8,584 microblogs form the

[2] https://github.com/fxsjy/jieba
[3] http://ir.dlut.edu.cn
[4] http://tcci.ccf.org.cn/conference/2013/

NoisyData dataset. In other words, a further 17.78% (8584/48,305) data are screened out.

3. Corpus Analysis

Through the above steps, the final cleaned emotion corpus with 39,661 microblogs consists of three parts: Part1, Part2 and Part3. The first two parts are obtained automatically and the third part manually. The distribution of different parts is shown in **Table 2.** Note that automatically obtained data accounts for more than 54.1% in the final selected corpus, which is translated to the reduction of manual work by 44.5% out of the D0 set (e.g. (Part1+Part2)/D0).

	Part1	Part2	Part3	Total
Size	14,197	7,228	18,236	39,721
Percentage (%)	35.74	18.35	45.91	100.00

Table 2 Obtained Corpus Distribution

The distribution of emotion classes is shown in **Table 3**, which shows that “sadness” and “happiness” have more samples whereas “surprise” and “fear” are much less. This is consistent with the manually annotated corpus NLP&CC2013 and, again, it shows an intrinsic data imbalance problem for emotion analysis.

Emotion	Number	Percentage (%)
Like	4540	11.45
happiness	9959	25.11
sadness	14052	35.43
anger	4562	11.50
disgust	4876	12.29
fear	661	1.67
surprise	1011	2.55
sum	39661	100.0

Table 3 Emotion Distribution

3.1 Selected Label Quality Analysis

To analyze the quality of the emotion corpus, we randomly sample about 5% of the data in each part with label balance control and ask another trained annotator to manually annotate the data based on the same rules from section 2.5. We compare the annotated label with the selected natural label and calculate the Kappa value. The result shown in **Table 4** indicates that the Kappa value achieves 0.941, 0.926 and 0.812 for lexicon, SVM and manual annotation based selection, respectively. This indicates that the label quality is quite high compared with the Kappa value of 0.713 of NLP&CC2013. The relatively high Kappa values indicate that proposed method is quite effective in obtaining quality data. For those text with inconsistent labels, we manually analyze the data and discover that the selected natural labels are more reasonable than manual ones. For example, in the sentence “今天放假 了，我会想念你们的！ (Holiday begins today, and I will miss you!)”, the original and the lexicon based label are both “sadness” whereas the manual label is “like”. However, the author feels sad because he will leave someone because of holiday and the lexicon-based method set it as sadness because of the word “miss”. But, the annotator allocates “like” because of “holiday”. The “sadness” label is more reasonable, which indicates the selected natural label is more reliable than the manual label.

Data	Size	Sample Size	Kappa
Part1	14,197	700	0.941
Part2	7,228	400	0.926
Part3	18,236	900	0.812

Table 4 Kappa Value of Automatically Selected Label

To prove that the acquired data is a useful resource, we compare the quality of the obtained corpus with the manual NLP&CC2013 corpus by training classifiers using them and test the classifiers on the NLP&CC2013 testing dataset. Given the same classifier, the assumption is that if the quality of a corpus is higher, the performance of the classifier trained on it should be better. We also compare the result with the classifier trained on the original hashtag dataset without selection (D0 dataset) and the NLP&CC2013 training dataset. For the NLP&CC2013 dataset, about half of them are with label “none”, which cannot be obtained through the hashtag, so we discard the “none” label both in NLP&CC2013’s training and testing data. The features are simple bag-of-words frequency. The classifier is Liblinear with the default parameter and metric is macro precision, recall and F-score, which can be calculated as follows:

$$\mathrm{Macro\_Precision} = \frac{1}{n}\sum_{\mathrm{i}} \frac{\mathrm{\#system\_correct(emotion{=}i)}}{\mathrm{\#system\_proposed(emotion{=}i)}}$$

$$Macro\_Recall = \frac{1}{n}\sum_{\mathrm{i}} \frac{\mathrm{\#system\_correct(emotion{=}i)}}{\mathrm{\#gold(emotion{=}i)}}$$

$$Macro\_F - score = \frac{2 \times \mathrm{Macro\_Precision} \times Macro\_Recall}{\mathrm{Macro\_Precision} + Macro\_Recall}$$

where *n* is the number of emotion labels, *#gold(emotion=i)* is the number of samples whose gold emotion label is *i*, *#system_correct(emotion=i)* is the number of samples whose predicted label is the same as gold label *i*, *#system_proposed(emotion=i)* is the number of samples whose predicted label is *i*.

The result is shown in **Table 5**, where the “Selected” is the built emotion corpus, the “Original” is the D0 dataset after Step 2. By using the selected data, the precision, recall and F-score improve over D0 by 11.1%, 5.0% and 7.6% respectively. The classifier trained on selected hashtag dataset achieves comparable result with the manually annotated NLP&CC2013 dataset. However, the recall the selected lexicon is lower. The reason is that the hashtag data is sampled from microblogs containing the emotional hashtags, which has a bias towards emotional data while the NLP&CC2013 comes from the whole microblogs that also contain non-emotional microblogs and the NLP&CC2013 training and testing data are consistent. This reveals one potential data bias problem of hashtag based selection method. When combining with non-hashtag microblogs, this problem may be solved, which will be our future work. This experiment shows the effectiveness of our proposed method for semi-automatic construction of emotion corpus from the raw data that contains noisy natural labels. This method can also be extended to other kinds of corpus construction.

Training data	Macro Precision	Macro Recall	Macro F-score
Selected	**0.4301**	0.3174	**0.3652**
Original	0.3870	0.3024	0.3395
NLP&CC2013 Training	0.3734	**0.3534**	0.3631

Table 5 Performance on NLP&CC2013 test dataset

3.2 Noisy Data Analysis

Now we focus on analyzing the 8,584 microblogs in **NoisyData** after Step 4 to evaluate the noise level in the natural label. This inconsistency results from two factors: (1). natural labels in **NoisyData** contain more noise, and (2). manual annotation is difficult, leading to incorrect manual labels. We denote the natural label as L1, the manual label as L2.

The basic idea to examine whether the natural label is noisy is to ask another trained annotator to annotate the **NoisyData** based on the same annotation rules in Section 2.5, and the labels are denoted as L3. Then we employ the voting strategy to determine the final label L. If L1 equals to L3, we denote it as a high-quality label, a noisy label otherwise. The statistic information is shown in **Table 6**, where the entries represent the percentage of L1=L3, L2=L3 and others, respectively. The total noise labels account for 72.5% in **NoisyData**, which converts to 12. 9% in the raw data (D0) after simple rule-based filtering. This indicates that about 12.9% of microblogs after simple rule based pre-processing contain noise.

	L1 = L3	L2=L3	Others
Percentage (%)	27.47	44.27	28.27

Table 6 Noisy Labels Distribution

4. Conclusion

In this paper, we present a method that combines automatic selection and manual annotation selection method based on natural hashtag labels to construct an emotion analysis corpus. Experiments show that this method can reduce manual annotation work and obtain high-quality corpus. Currently, 39,661 Chinese microblogs with high-quality emotion labels are obtained and we make it public available[5]. In addition, there is one potential issue is that such hashtag based corpus has data bias problem that the obtained data has no “none” label data. In future work, we will explore methods to solve the problem and explore the usage of this corpus on emotion analysis tasks in the future.

5. Acknowledgements

We give our thanks to the anonymous reviewers for the helpful comments. The work presented in this paper is supported by Hong Kong Polytechnic University (PolyU RTVU and CERG PolyU 15211/14E) and the National Nature Science Foundation of China (project number: 6127229).

6. References

Alm, Ebba Cecilia Ovesdotter. 2009. *Affect in Text and Speech*. VDM Verlag Dr. Müller.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning Research* 9: 1871–74.

Mohammad, Saif M, and Svetlana Kiritchenko. 2014. Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*.

Quan, Changqin, and Fuji Ren. 2010. A Blog Emotion Corpus for Emotional Expression Analysis in Chinese. *Computer Speech & Language* 24 (4): 726–49.

Scherer, K. R., and H. G. Wallbott. 1994. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *Journal of Personality and Social Psychology* 66 (2): 310–28.

Strapparava, Carlo, and Rada Mihalcea. 2007. Semeval-2007 Task 14: Affective Text. Association for Computational Linguistics.

Suykens, Johan AK, and Joos Vandewalle. 1999. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* 9 (3): 293–300.

Tang, Duyu, Bing Qin, Ting Liu, and Zhenghua Li. 2013. Learning Sentence Representation for Emotion Classification on Microblogs. In *Natural Language Processing and Chinese Computing*, 212–23. Springer.

Wang, Wenbo, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing Twitter ‘Big Data’ for Automatic Emotion Identification. In *Privacy, Security, Risk and Trust (PASSAT), International Conference on Social Computing (SocialCom)*, 587–92.

Xu, Linhong, Hongfei Lin, Pan Yu, Hui Ren, and Jianmei Chen. 2008. Constructing the Affective Lexicon Ontology [J]. *Journal of the China Society for Scientific and Technical Information* 2: 006.

Yuanlin, YAO, WANG Shuwei, XU Ruifeng, LIU Bin, GUI Lin, LU Qin, and WANG Xiaolong. 2014. The Construction of an Emotion Annotated Corpus on Microblog Text. *Journal of Chinese Information Processing* 28 (5): 83–91.

[5] https://github.com/MingleiLI/emotion_corpus_weibo