A Corpus-based Multi-label Emotion Classification using Maximum Entropy

Ye Wu[1,2] and Fuji Ren[1,2]

[1] Faculty of Engineering, The University of Tokushima, Tokushima, Japan
[2] School of Computer, Beijing University of Posts and Telecommunications, Beijing, China

Abstract. Thanks to the Internet, blog posts online have emerged as a new grass-roots medium, which create a huge resource of text-based emotion. Comparing to other ideal experimental settings, what we obtained from the World Wide Web evolve and respond more to real-world events. In this paper, our corpus consists of a collection of blog posts, which annotated as multi-label to make the classification of emotion more precise than the single-label set of basic emotions. Employing a maximum entropy classifier, our results show that the emotions can be clearly separated by the proposed method. Additionally, we show that the micro-average F1-scores of multi-label detection increase when the mount of available training data further increases.

1 Introduction

From the informative contents involved in text, not only communication information, but also private attitudes such as emotion states can be found. Human-computer interaction would be more natural and social if we can capture the emotion information occurring in texts. Researchers have tried to detect the emotion state of the intelligent interfaces users in many ways, and these works are dubbed as “affective computing” by Picard [1], which have recognized the potential and importance of affect to human-computer interaction.

Some approaches to text-based emotion classification demonstrated good performance. Liu et al. did a rare study in text-based emotional prediction, which used a database of common-sense knowledge and created affect models to form a representation of the emotional affinity of a sentence [2]. Work on emotion classification in the view of text-to-speech was carried out by Cecilia [3]. They addressed text-based emotion prediction as the first task in text-to-speech synthesis. Mihalcea deployed a series of interesting experiments to find happiness [4], whose study is based on a collection of blog posts from LiveJournal.com blog community.

As Nass study suggests, people most naturally interact with their computer in a social and affectively meaningful way just like with other people [5]. The authors of previous work almost adopt the notion of basic emotions [6], thus using six emotion categories: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. And in some other works the authors just recognized two emotional passages (i.e. positive versus negative, or happy versus sad). In our experiments, we took Chinese blog posts as the sources of our corpus, and when we showed some text samples to testers, no one of them thought

Wu Y. and Ren F. (2009).
A Corpus-based Multi-label Emotion Classification using Maximum Entropy.
In *Proceedings of the 6th International Workshop on Natural Language Processing and Cognitive Science* , pages 103-110
DOI: 10.5220/0002169901030110

employing one of the six basic emotions can describe the emotion information to mass of sentences precisely because human is more sensitive. Therefore previous models mentioned above usually result in an emotion distortion problem, which means that the results of the classification can not denote emotion states correctly. On the other hand, Mishne experiment with emotion classification using cross-training and an ontology of over 100 moods shows rather low accuracy even when the training corpus is very large [7]. So we proposed that maybe we should consider this problem as similar with RGB model. Thus at first we could complete the task of automatically classifying texts into categories according to descriptions of their emotions, and then find the co-occurrence patterns and their valence between labels, and finally synthesize those labels assigned to one text to a more detailed and precise one described in Chinese, depending on the rules learned from the previous step. In this paper, we utilized multi-label classification that has been applied on topic detection and some other domains, to perform experiments as the first single step of the work we focus on.

The remaining parts of this paper are organized as follows. Section 2 provides a brief introduction of the blog emotion corpus we used. Section 3 describes the model of our method, including the representation of the training samples, and the approach to multi-label classification we employed. In Section 4 we present the results of our experiments, compared using different evaluation measures. Finally, Section 5 describes with a discussion and Section 6 concludes on this investigation and presents the possible directions for the future work.

2 A Blog Emotion Corpus

With the increase in the web's accessibility in the last years, the amount of weblogs has risen dramatically and the so-called blogsphere attracts new research interests.

In this paper, our study is based on a collection of blog posts from various Chinese blog communities, because these sources for modeling are recognized as more private, honest, and polemic than opinions voiced in other style [4]. Emotion recognition models proposed before were usually implemented on a roughly annotated corpus, so ambiguity would be caused by the emotion corpus itself. Some corpora with which the previous related work experimented are actually emotionless. Lee et al. reported a high accuracy of emotion recognition, but 73.17% of their corpus are neutral [8], and thus they obtained a baseline system with the accuracy of 73.17%, which can be achieved easy by simply marking the neutral texts. Our current corpus Ren-CECps (Chinese Emotion Corpus of RenLab) 1.0 is manually annotated as a sentence level one consists of 198 documents, 5608 sentences, and 135,606 Chinese characters. Annotators worked in triples on the same texts. They have been trained and work independently in order to avoid any annotation bias and get a true understanding of the task difficulty. The goal of our annotation project is to annotate a corpus of approximately 1000 blog posts, and the details of the corpus would be published after completing it. Comparing the six basic categories of emotion, each annotator marks the sentences with one or more of eight emotions listed as follow: **Anger, Anxiety, Expect, Hate, Joy, Love, Sorrow and Surprise**. In order to find the rules how each emotion keyword or phrase correspond to the labels in our future work, the image value of every emotion term attributed to every emotion label is also

scored by our annotators. Other than arbitrary keywords with parts of speech tags which were usually utilized in previous works, we mark emotion keywords without POS tags, for the flexibility and diversity of part of speech in Chinese. Besides, our emotion terms include long structural emotion phrases that are accustomed to express person's affect in Chinese, and special emotion phrases like *idioms*, *proverbs*, and *words of separation and reunion* are also annotated in our corpus.

Furthermore, advanced linguistic and practical features could be considered in the task of emotion classification, such as degree word, negative word, conjunction, rhetoric, and punctuation are also labeled detailedly in our sentence-level annotation. Relative experiments will be carried on in the next steps of our work.

3 The Model for Emotion Classification

Feature-based statistical classification is applied to text emotion recognition, since basic algorithm within emotion recognizers is very similar to text categorization and topic detection. This part we employ maximum entropy classifier as our machine learning method to the automatic emotion classification, and we extract feature sets from the corpus we constructed.

3.1 Text Representation

Maximum entropy (ME or maxent for short) is a general purpose machine learning model that has been successfully applied in Natural Language Processing (NLP)[9-11]. This conditional model for text classification provides a rich framework for representing relationships between classes and features of a given domain, with the goal of best accounting for some training data. Given training data D = $\{\langle \mathbf{x}_1, y_1\rangle, \langle \mathbf{x}_2, y_2\rangle, \ldots, \langle \mathbf{x}_r, y_r\rangle\}$ (having vocabulary *V* and set *Y* of classes), any real-valued function of the text x and class y, f_i(x,y) can be treated as a feature.

As we have annotated a blog emotion corpus including an abundance of emotional contents, we take a part of it to perform our experiments at sentence level. There are 5608 sentences in the corpus we chose. Firstly we excluded the neutral sentences without any emotion information at all, and the remaining 4968 sentences with emotion labels are utilized in the later study.

We represent text feature vectors in several different ways, three utilzing correlative information between terms and multiple labels, and two just considering single label.Then we compare their performance. In multi-label situation, an element in the feature vector can be described as follow. Given sentence x with label vector **y**, if the sentence was labeled with y_k involved in **y**,

$$f_{w_k,y_k}(x,\mathbf{y}) = V(x,w_k),$$

otherwise,

$$f_{w_k,y_k}(x,\mathbf{y}) = 0,$$

here feature value $V(x,w_k)$ equals to the summation of values marked manually by annotators for term w_k in x, the number of occurrences of term w_k in x, and the presence

of term w_k in x not considering the frequency respectively. According to these definitions we can obtain three feature sets: correlative labels value feature set, correlative labels frequency feature set and correlative labels present feature set. Accordingly, in single-label situation, if the sentence was labeled with y,

$$f_{w_k,y}(x,y) = V(x,w_k),$$

otherwise,

$$f_{w_k,y}(x,y) = 0,$$

here feature value $V(x,w_k)$ equals to the number of occurrences of term w_k in x, and the presence of term w_k in x not considering the frequency respectively. According to these definitions we can obtain two feature sets: frequency feature set and present feature set. Note that the single-label situation obviates manually marked values because every keywords and phrases in our corpus are annotated in multi-label way. We would see multi-label feature representation for classification resulted in better performance in the next section.

3.2 Binary Pruning

Binary pruning is one common approach to automatic classification of data belonging to more than one class by first training an independent maximum entropy binary classifier for each label. For each sentence, each classifier has a probability score for the respective class, and then classifies a sentence into a category if the corresponding binary classifier scores above some threshold. Comparative experiments using 14 classifiers have been conducted Yang [12]. This approach is considered to be unable to exploit dependencies between labels [13], but the advantage of this technique is that it is possible to correctly classify test texts whose actual combinations do not occur in the training data, so it is more feasible. Furthermore, by offering a hierarchical solution selecting only the best scoring classes for each text, this approach reduces overfitting of the model parameters to the training data.

4 Experiments and Evaluations

Since each sentence from the corpus is represented as a feature vector in the ways we referred in Section 3, we train maximum entropy classifiers and present the results from experiments in this section. To test our approach, we did two sets of experiments and compared the results with several widely used metrics.

4.1 Emotion Classification

In the first experiment, accuracy of every maximum entropy binary classifier that classify sentences either **with** or **without** a special kind of emotion, using 5-fold cross validation with all 4968 sentences having emotion labels, and average accuracy for the five different feature sets, are included in Table 1, in which c& v, c& f, c& p, fre, and

Table 1. Classification accuracy for 8 emotions, 5 feature sets, given as percents.

Emotion	c&v	c&f	c&p	fre	pre
Anger	85.58	85.52	85.56	83.32	83.02
Anxiety	76.09	76.21	75.89	71.09	70.74
Expect	74.94	74.36	75.13	69.83	70.29
Hate	76.94	76.66	77.06	74.16	73.89
Joy	70.27	70.51	69.93	68.62	68.22
Love	63.80	63.68	64.41	62.88	62.80
Sorrow	77.14	76.70	76.59	73.98	74.22
Surprise	90.03	89.89	89.81	89.03	89.16
Average	76.85	76.69	76.80	74.11	74.04

pre stand for correlative labels value feature set, correlative labels frequency feature set, correlative labels present feature set, frequency feature set and present feature set respectively.

The average accuracy for each of all five feature sets is measured above 70%, and the best result is measured at 76.85%. As it turns out, the labels in this corpus are clearly separated, and therefore we can use this data to learn the co-occurrence of multiple labels indicated in the blog sentences. The numbers in this table also show that the results using correlative labels feature sets have about 3% improvement over single label feature sets. In the results for three correlative labels feature sets, the set with manually marked image value shows best, and binary present features outperform frequency features, which is also inferred in related tasks using other classifiers [14].

4.2 Multi-label Experimental Evaluations

The common notions of precision and recall have been applied to measure performance of a statistical classification. This set of experiments are built with different training sets and evaluated with the same test set consisting of 1000 sentences.

Table 2 shows the results for each classifier on each emotion class. Here we trained with the correlative labels value feature set which performed the best at average accuracy from 5-fold cross validation in section 4.1, with 1000 to 3968 training samples (1000/ 2000/ 3000/ 3968 samples).

Overall performance for multi-label classification is measured by summing up the values over all classes and doing a division. Here TP_i (true positive) is the number of documents correctly assigned to emotion class E_i (the number of classes is $|E|$), and FP_i (false positives), FN_i (false negatives) and TN_i (true negative) are defined accordingly. We calculate precision and recall according to the micro-average approach [12]:

$$Precision_{micro} = \frac{\sum_{i=1}^{|E|} TP_i}{\sum_{i=1}^{|E|}(TP_i + FP_i)} \quad (1)$$

$$Recall_{micro} = \frac{\sum_{i=1}^{|E|} TP_i}{\sum_{i=1}^{|E|}(TP_i + FN_i)} \quad (2)$$

Table 2. Results of binary classifiers using 1000, 2000, 3000, 3968 training samples.

Samples	Emotion	Precision	recall	F1-score
1000	Anger	12.77	57.75	20.92
	Anxiety	38.10	32.00	34.78
	Expect	49.29	34.67	40.70
	Hate	20.55	50.34	29.18
	Joy	55.71	30.23	39.20
	Love	63.41	36.28	46.15
	Sorrow	47.17	21.28	29.33
	Surprise	52.94	25.00	33.96
2000	Anger	15.38	50.70	23.61
	Anxiety	43.42	33.00	37.50
	Expect	56.05	41.67	47.80
	Hate	22.88	41.61	29.52
	Joy	56.41	34.11	42.51
	Love	61.23	39.30	47.88
	Sorrow	44.93	26.38	33.24
	Surprise	47.73	29.17	36.21
3000	Anger	18.18	50.70	26.77
	Anxiety	41.51	33.00	36.77
	Expect	56.52	39.00	46.15
	Hate	23.87	35.57	28.57
	Joy	52.97	44.96	48.64
	Love	60.92	46.05	52.45
	Sorrow	49.46	38.72	43.44
	Surprise	33.33	31.94	32.62
3968	Anger	19.23	49.30	27.67
	Anxiety	38.42	34.00	36.07
	Expect	53.21	38.67	44.79
	Hate	25.12	36.24	29.67
	Joy	56.70	49.22	52.70
	Love	62.87	48.84	54.97
	Sorrow	49.77	45.11	47.32
	Surprise	34.57	38.89	36.60

$$F1_{micro} = \frac{2 \cdot Precision_{micro} \cdot Recall_{micro}}{(Precision_{micro} + Recall_{micro})} \tag{3}$$

Table 3 depicts the macro-average F1-scores as well as micro-average F1-scores with 1000, 2000, 3000 and 3968 training, using binary pruning. The macro-average of F1 scores is the mean of the F1-scores of all the labels, attributing equal weights to the label F1 scores, while the micro-average is the F1-score obtained from the summation of contingency matrices for all binary classifiers. Thus the micro-average metric gives equal weight to all classifications, so that F1 scores of larger classes influence the metric more than F1 scores of smaller classes. Comparing micro-average and macro-average of F1-scores for all emotion class labels facilitates evaluation of the performance of multi-label classifiers. Mean accuracy over all emotion classes for different training sets are also included in Table 3.

Table 3. Results of multi-label classification using binary pruning, with 4 training sets.

Measure	1000 samples	2000 samples	3000 samples	3968 samples
macro-F1	34.28	37.28	39.43	41.22
micro-F1	35.45	36.52	42.45	44.30
mean accuracy	73.33	75.61	76.27	76.61

5 Discussion and Conclusion

The classification performance of binary classifiers which use correlative information involved feature sets we defined, has about 3% improvement. Nevertheless, observations of multi-label micro-average F1-scores look not so good. One of possible reasons is the current data set appears to be not enough to gather meaningful statistics. As we worked on small text units - sentences, some of them create few features in the feature vector, making training data very sparse. The results of micro-average precision, recall and F1-score indicate that the performance of our method is still improving with the increment of the amount of samples in training set.

The results also show that different kind of emotion follow different track with the increment of training data. **Love** is the largest class in all of eight emotion classes while **Surprise** is the smallest one. Not surprisingly, the result of smaller class is worse than the result of larger class according to Table 2 in the last section. When the maximum entropy classifier for the small emotion class begins to train, this emotion is absent in the most of labels. Another observation that should be noted in Table 2 is comparing positive emotions such as **Love** and **Joy**, classifiers for their opposite emotions like **Hate** and **Sorrow** performed not so well. And the F1-scores of **Anxiety** and **Expect** declined on the contrary when the amount of training samples increases. We think that we should conduct some investigations to explain these phenomena. The degressive F1-scores of emotion class **Anxiety** and **Expect** is most likely due to the borderline between these two kinds of emotions is ambiguous, and sometimes **Expect** also be regarded as positive **Anxiety**. Classifiers for each emotion should be discussed separately in more detail considering their characteristics.

6 Conclusions

In this paper we have presented preliminary experiments of text-based multi-label classification using a blog emotion corpus, with finding emotional information from self-expression blog posts as our goal. Using the correlative labels value feature set to create feature vectors, the reported mean accuracy and micro-average of F1-scores make we believe that our results can still improve by increasing the training set size. Additionally, our classification accuracy is not substantially worse than annotation task because of the subjective error.

In future work, we wish to take more advanced linguistic features into account, utilizing all the useful information marked in the blog emotion corpus, after our current annotation project is completed. And more complex models for multi-label classification would be tried then. In the processing of statistical learning, it is hoped that the

co-occurrence patterns and their valence between labels could be found to eliminate the emotion distortion, and help affect understanding not only for human-computer interaction but also psychological research.

Acknowledgements

We are grateful to all the annotators. This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 19300029.

References

1. Rosalind W. Picard: *Affective Computing*. The MIT Press, Mass. (1997).
2. Hugo Liu, Henry Lieberman, Ted Selker: *A Model of Textual Affect sensing using real-world knowledge*. ACM Conference on Intelligent User Interfaces (2003) 125–132.
3. Cecilia Ovesdotter Alm, Dan Roth, Richard Sproat: *Emotions from Text: Machine Learning for text-based emotion prediction*. HLT/EMNLP (2005).
4. Rada Mihalcea, Hugo Liu: *A corpus-based approach to finding happiness. Proceedings of Computational approaches for analysis of weblogs*, AAAI Spring Symposium (2006).
5. Clifford Nass, Jonathan Steuer, Ellen Tauber: *Computers are Social Actors*. Proceedings of CHI'94 (1994) 72–78.
6. Paul Ekman: Facial expression and emotion. American Psychologist, Vol. 48 (1993) 384–392.
7. Gilad Mishne: *Experiments with Mood Classification in blog posts*. Proceedings of the 1st Workshop on Stylistic Analysis of Text for Information Access, Brazile (2005).
8. Cheongjae Lee, Gary Geunbae Lee: *Emotion Recognition for Affective user Interfaces using Natural Language Dialogs*. 16th IEEE International Conference on Robot & Human Interactive Communication, Korea (2007).
9. Adam Berger, Stephen Della Pietra, Vincent Della Pietra: *A maximum Entropy approach to Natural Language Processing*. Computational Linguistics, Vol. 22 (1996).
10. Stephen Della Pietra, Vincent Della Pietra, John Lafferty: *Inducing Features of Random Fields*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19 (1997).
11. Kamal Nigam, John Lafferty, Andrew McCallum: *Using maximum entropy for text classification*. IJCAI-99 Workshop on Machine learning for Information Filtering (1999) 61–67.
12. Yiming Yang: *An Evaluation of Statistical Approaches to Text Categorization.* Journal of Information Retrieval(1998).
13. Nadia Ghamrawi, Andrew McCallum: *Collective Multi-label Classification*. Proceedings of the 3005 ACM Conference on Information and Knowledge Management (2005) 195–200.
14. Bo Pang, Lillian Lee: A sentimental education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. Proceedings of ACL (2004) 217–278.