Predicting Emotion Labels for Chinese Microblog Texts

Zheng Yuan[1], Matthew Purver[2]

School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS
[1]yuanzheng.liliian@hotmail.com
[2]m.purver@qmul.ac.uk

Abstract. We describe an experiment into detecting emotions in texts on the Chinese microblog service Sina Weibo using distant supervision with various author-supplied conventional labels (emoticons and smilies). Existing word segmentation tools proved unreliable; better accuracy was achieved using character-based features. Accuracy varied according to emotion and labelling convention: while smilies are used more often, emoticons are more reliable. Happiness is the most accurately predicted emotion (85.9%). This approach works well and achieves 80% accuracies for "happy" and "fear", even though the performances for the seven emotion classes are quite different.

Keywords: Social Media, Sina Weibo, Emotion Detection, Emoticons, Smilies, Distant Supervision, N-gram lexical features

1   Introduction

Social media has become a very popular communication tool among Internet users. Sina Weibo (hereafter Weibo), is a Chinese microblog website. Most people take it as the Chinese version of Twitter; it is one of the most popular sites in China, in use by well over 30% of Internet users, with a similar market penetration that Twitter has established in the USA (Rapoza, 2011 [1]), and has therefore become a valuable source of people's opinions and sentiments.

Microblog texts (statuses) are very different from general newspaper or web text. Weibo statuses are shorter and more casual; many topics are discussed, with less coherence between texts. Combining this with the huge amount of lexical and syntactic variety (misspelt words, new words, emoticons, unconventional sentence structures) in Weibo data, many existing methods for emotion and sentiment detection which depend on grammar- or lexicon-based information are no longer suitable.

Machine learning via supervised classification, on the other hand, is robust to such variety but usually requires hand-labeled training data. This is difficult and time-consuming with large datasets, and can be unreliable when attempting to infer an author's emotional state from short texts (see e.g. Purver & Battersby, 2012 [2]). Our solution is to use distant supervision: we adapt the approach of (Go et al., 2009 [3]; Purver & Battersby, 2012 [2]) to Weibo data, using emoticons and Weibo's built-in

smilies as author-generated emotion labels, allowing us to produce an automatic classifier to classify Weibo statuses into different basic emotion classes. Adapting this approach to Chinese data poses several research problems: finding accurate and reliable labels to use, segmenting Chinese text and extracting sensible lexical features.

Our experiments show that choice of labels has a significant effect, with emoticons generally providing higher accuracy than Weibo's smilies, and that choice of text segmentation method is crucial, with current word segmentation tools providing poor accuracy on microblog text and character-based features proving superior.

2 Background

2.1 Sentiment/Emotion Analysis

Most research in this area focuses on sentiment analysis – classifying text as positive or negative (Pang and Lee, 2008 [4]). However, finer-grained emotion detection is required to provide cues for further human-computer interaction, and is critical for the development of intelligent interfaces. It is hard to reach a consensus on how the basic emotions should be categorised, but here we follow (Chuang and Wu, 2004 [5]) and others in using (Ekman, 1972 [6])'s definition, providing six basic emotions: anger, disgust, fear, happiness, sadness, surprise.

2.2 Distant Supervision

Distant supervision is a semi-supervised learning algorithm that combines supervised classification with a weakly labeled training dataset. (Go et al., 2009 [3]) and (Pak and Paroubek, 2010 [7]), following (Read, 2005 [8]), use emoticons to provide these labels to classify positive/negative sentiment in Twitter messages with above 80% accuracy.

(Yuasa et al., 2006 [9]) showed that emoticons have an important role in emphasizing the emotions conveyed in a sentence; they can therefore give us direct access to the authors' own emotions. (Purver and Battersby, 2012 [2]) thus used a broader set of emoticons to extend the distant supervision approach to six-way emotion classification in English, and we apply a similar approach. However, in addition to the widely used, domain-independent emoticons, other markers have emerged for particular interfaces or domains. Sina Weibo provides a built-in set of smilies that can work as special emoticons that help us better understand authors' emotions.

2.3 Chinese Text Processing

In Chinese text, sentences are represented as strings of Chinese characters without explicit word delimiters as used in English (e.g. white space). Therefore, it is important to determine word boundaries before running any word-based linguistic processing on Chinese. There is a large body of research into Chinese word segmentation (Fan and Tsai, 1988 [10]; Sproat and Shih, 1990 [11]; Gan et al, 1996 [12]; Guo, 1997

[13]; Jin and Chen, 1998 [14]; Wu, 2003 [15]). Among them, the basic technique for identifying distinct words is based on the lexicon-based identification scheme (Chen and Liu, 1992). This approach performs word segmentation process using matching algorithms: matching input character strings with a known lexicon. However, since the real-world lexicon is open-ended, new words are coming out every day – and this is especially true with social media. A lexicon is therefore difficult to construct or maintain accurately for such a domain.

3   Data

3.1   Corpus Collection

Our training data consisted of Weibo statuses with emoticons and smilies. Since Weibo has a public API, training data can be obtained through automated means. We wrote a script which requested the statuses public_timeline API[1] every two minutes and inserted the collected data into a MySQL database. We collected a corpus of Weibo data, filtering out messages not containing emotion labels (see below and Table 2 for details).

3.2   Emotion Labels

We used two kinds of emotion labels (emoticons and smilies) as our noisy labels. The emoticons and smilies are noisy themselves: ambiguous or vague. Not all the emoticons and smilies have close relationships with the emotion classes. And some emoticons and smilies may be used in different situations, as different people have different understandings. Emoticons here are Eastern-style emoticons, very different from Western-style ones (see e.g. Kayan et al., 2006 [16]). Smilies are Sina Weibo's built-in smilies. Initial investigation found that not all emoticons and smilies can be classified into Ekman's six emotion classes; and for some lesser used labels, authors have widely different understandings. We identified the most widely used and well-known emoticons/smilies to use as labels – see Table 1.

Table 1. Conventional markers used for emotion classes

Emotion	Emoticons	Smilies
surprise	OMG; (0.o); (O_o); (@_@); (O_O); (O?O)	[吃惊 chi-jing "surprise"]
disgust	N/A	[吐 tu "sick"]
happy	(^_^); (*^__^*);(^o^); (^.^);O(∩_∩)O;	[嘻嘻 xi-xi "heehee"];
		[哈哈 ha-ha "haha"];
		[鼓掌 gu-zhang "applaud"];

angry	╭(╯＾╰)╮ ; (╰_╯)	😊[太开心 da-kai-xin "so happy"] 😡[怒 nu "anger"]; 😠[怒骂 nu-ma "curse"]; 😤[哼 heng "humph"]; 😒[鄙视 bi-shi "disdain"]
fear	Just use the keyword 害怕 hai-pai "fear"	
sad	(T_T); (T.T); (T^T); (ㅠ.ㅠ); (╯＿╰); (︶︿︶);	😭[泪 lei "tear"]; 😲[失望 shi-wang "disappointed"]; 🙁[悲伤 bei-shang "sad"]

3.3 Text Processing

We used a Chinese language selection filter to filter out all other language characters or words, removed URLs, Weibo usernames (starting with @), digits, and any other notations, e.g., *, ¥, only leaving Chinese characters. We then removed the emoticons and smilies from the texts, replacing them with positive/negative labels for the relevant emotion classes for training and testing purposes. We then extracted different kinds of lexical features: segmented Chinese words, Chinese characters, and higher order n-grams.

For word-based features, we need to segment the sentences. There are lots of Chinese word segmentation tools; however, many are unsuitable for online social media text; we chose pymmseg[2], smallseg[3] and the Stanford Chinese Word Segmenter[4], which all appeared to give reasonable results. Pymmseg uses the MMSEG algorithm (Tsai, 2000 [17]). Smallseg is an open sourced Chinese segmentation tool based on DFA. The Stanford Segmenter is CRF-based (Tseng et al, 2005 [18]).

3.4 Corpus Analysis

Our database contains 229,062 Weibo statuses with emotion labels; Table 2 shows statistics. The number of Weibo statuses varied with the popularity of the labels themselves: "happy" and "sad" labels are much more frequent than others; very similar results are observed in English Twitter statuses (see e.g. [2]), suggesting that these frequencies are relatively stable across very different languages.

Table 2. Number of statuses per emotion class

Emotion	Mixed	Emoticons	Smilies

---

surprise	347	63	284
disgust	142	N/A	142
happy	5685	712	4973
angry	2318	9	2305
fear	480	Key words: 480	
sad	5422	1064	4358

Overall frequencies show that users of Weibo are more likely to use the built-in smilies rather than emoticons. One possible reason is that smilies can be inserted with a single mouse click, whereas emoticons must be typed using several keystrokes – Eastern-style emoticons are usually made of five or more characters.

4   Experiments and Discussions

Classification was using support vector machines (SVMs) (Vapnik, 1995 [19]) throughout, with the help of the LibSVM tools (Chang and Lin, 2001 [20]). The performance was evaluated using 10-fold cross validation. Our datasets were balanced: a dataset of size N contained N/2 positive instances (statuses containing labels for this emotion class) and N/2 negative ones (statuses containing labels from other classes). For the N/2 negative instances, we randomly selected instances from other emotion classes for larger datasets (N>1200), but ensured an even weighting across negative classes for smaller sets to prevent bias towards one negative class. Because of the different frequency of different emotion labels, we mainly focused on "happy", "angry" and "sad", and present tentative results for the other emotion classes.

4.1   Segmented Words-VS-Characters

In the first experiment, we investigated the effect of different segmentation tools and compared word-based vs character-based features.

After testing on "angry", "happy" and "sad", we found that pymmseg outperformed the other tools; we therefore used pymmseg for later experiments. However, as we increased the dataset size, we found that character-based features had even better performance than word features (using pymmseg) for all three classes. Our results suggest that we could just use Chinese characters, rather than doing any word segmentation - see Figure 1.

Examination of the segmented data showed that the segmentation tools didn't work well with our social media data and made lots of mistakes. In addition, all segmentation tools produced many segmented words which were actually just one character. The use of character-based features was therefore preferred.
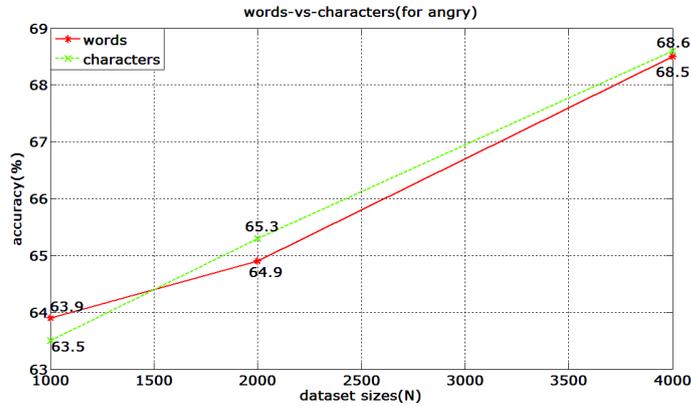
Fig. 1. Words-vs-Characters (for "angry")

4.2 Increasing Accuracy

In the second experiment, we tried to improve the overall performance.

Whether higher-order n-grams are useful features appears to be a matter of some debate. (Pang et al., 2002 [21]) report that unigrams outperform bigrams when classifying movie reviews by sentiment polarity, but (Dave et al., 2003 [22]) find that bigrams and trigrams can give better product-review polarity classification.
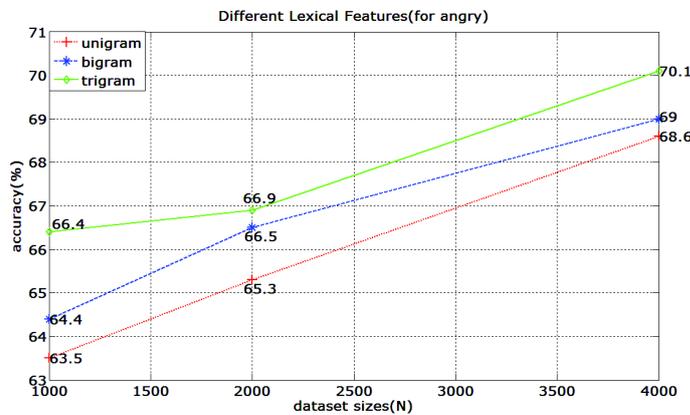

Fig. 2. Performance of n-grams (for "angry")

Results showed that higher-order n-grams are useful features for our wide-topic social media Weibo data. Bigrams and trigrams outperform unigrams for all these three emotion classes (see Figure 2). In our experiments with bigram and trigram features, we also included the lower-order n-grams (unigrams, bigrams), as there are lots of Chinese words with only one character. Our experiments also showed that increasing our dataset sizes increased accuracy; as our dataset sizes increase over time, we therefore expect improvements in accuracy (Figs 1 and 2).

45

Table 3. Best performance for three emotion classes

Emotion	Dataset Size N	Accuracy
angry	4000	70.1%
happy	8000	78.2%
sad	8000	69.6%

4.3 Smilies-vs-Emoticons

Our last experiment compared the two different kinds of labels: emoticons and smilies.

Table 4. Results of emoticons-vs-smilies (N=1200)

Emotion	Mixed	Emoticons	Smilies
happy	73.8%	85.9%	74.6%
sad	62.8%	67.5%	66.0%

Results showed that the emoticon labels were easier to classify than smilies. By looking at the data, we found that people use emoticons in a more systematic or consistent way. They use emoticons to tell others what their real emotions are ("happy", "sad" etc.), but on the other hand, they use smilies for a much bigger range of things, such as jokes, sarcasm, etc. Some people use smilies just to make their Weibo statuses more interesting and lively, apparently without any subjective feelings.

5 Conclusion

We used SVMs for automatic emotion detection for Chinese microblog texts. Our results show that using emoticons and smilies as noisy labels is an effective way to perform distant supervision for Chinese. Emoticons seem to be more reliable for emotion detection than smilies. It was also found that, when dealing with social media data, many Chinese word segmentation tools do not work well. Instead, we can use characters as lexical features and performance improves with higher-order n-grams. Increasing the dataset size also improves performance, and our future work will examine larger sets.

References

1. Kenneth Rapoza: China's Weibos vs US's Twitter: And the Winner Is? http://www.forbes.com/sites/kenrapoza/2011/05/17/chinas-weibos-vs-uss-twitter-and-the-winner-is/ (2011)
2. Matthew Purver and Stuart Batersby: Experimenting with Distant Supervision for Emotion Classification. In: 13th Conference of the European Chapter of the Association for Computational Linguistics. (2012)
3. Alec Go, Richa Bhayani, and Lei Huang: Twitter sentiment classification using distant supervision. Master's thesis, Stanford University. (2009)

4. Bo Pang and Lillian Lee: Opinion mining and sentiment analysis. Foundations and Trends in In- formation Retrieval, 2(1–2):1–135. (2008)

5. Ze-Jing Chuang and Chung-Hsien Wu: Multi- modal emotion recognition from speech and text. In: Computational Linguistics and Chinese Language, 9(2):45–62. (2004)

6. Paul Ekman: Universals and cultural differ- ences in facial expressions of emotion. In J. Cole, editor, Nebraska Symposium on Motivation 1971, volume 19. University of Nebras- ka Press. (1972)

7. Alexander Pak and Patrick Paroubek: Twitter as a corpus for sentiment analysis and opin- ion mining. In: 7th Conference on International Language Resources and Evaluation. (2010)

8. Jonathon Read: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: 43rd Meeting of the Association for Computational Linguis- tics. (2005)

9. Masahide Yuasa, Keiichi Saito and Naoki Mukawa: Emoticons convey emotions without cognition of faces: an fMRI study. CHI EA '06. ISBN: 1-59593-298-4, doi: 10.1145/1125451.1125737 (2006)

10. Fan, C. K., & Tsai, W. H.: Automatic word identification in Chinese sentences by the re- laxation technique. Computer Processing of Chinese & Oriental Languages, 4, 33-56. (1988)

11. Richard Sproat and Chilin Shih: A Statistical Method for Finding Word Boundaries in Chinese Text. Computer Processing of Chinese and Oriental Languages, 4, 336-351, (1990)

12. Kok-Wee Gan, , Martha Palmer, and Kim-Teng Lua: A statistically emergent approach for language processing: Application to modeling context effects in ambiguous Chinese word boundary perception. Computational Linguistics, 22(4):531–53. (1996)

13. Jin Guo: Critical tokenization and its properties. Computational Linguistics, 23(4):569–596. (1997)

14. Wangying Jin, and Lei Chen: Identifying unknown words in Chinese corpora. In: First Workshop on Chinese Language, University of Pennsylvania, Philadelphia. (1998)

15. Andi Wu: Customizable segmentation of morphologically derived Words in Chinese. In: Computational Linguistics and Chinese Language. 8(2). (2003)

16. Shipra Kayan, Susan R. Fussell and Leslie D. Setlock: Cultural differences in the use of instant messaging in Asia and North America. In: 20th anniversary conference on Com- puter supported cooperative work, Banff, Alberta, Canada. (2006)

17. Chih-Hao Tsai: MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm. http://technology.chtsai.org/mmseg/ (2000)

18. Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning: A Conditional Random Field Word Segmenter. In: Fourth SIGHAN Workshop on Chinese Language Processing. (2005)

19. Vladimir N. Vapnik: The Nature of Statistical Learning Theory. (1995)

20. Chih-Chung Chang and Chih-Jen Lin: LIBSVM: a library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (2001)

21. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan: Thumbs up? Sentiment classification using machine learning techniques. In: Conference on Empirical Methods in Natural Language Processing, pages 79–86. (2002)

22. Kushal Dave, Steve Lawrence, and David M. Pennock: the peanut gallery: Opinion ex- traction and semantic classification of product reviews. In: WWW, pages 519–528. (2003)