

Predicting Emotion Labels for Chinese Microblog Texts

Zheng Yuan and Matthew Purver

Abstract We describe an experiment into detecting emotions in texts on the Chinese microblog service Sina Weibo (www.weibo.com) using distant supervision via various author-supplied emotion labels (emoticons and smilies). Existing word segmentation tools proved unreliable; better accuracy was achieved using character-based features. Higher-order n-grams proved to be useful features. Accuracy varied according to label and emotion: while smilies are used more often, emoticons are more reliable. Happiness is the most accurately predicted emotion, with accuracies around 90% on both distant and gold-standard labels. This approach works well and achieves high accuracies for happiness and anger, while it is less effective for sadness, surprise, disgust and fear, which are also difficult for human annotators to detect.

1 Introduction

Social media has become a very popular communication tool among Internet users. In China, the number of users of social networking websites had reached 288 million by the end of June 2013. The proportion of social networking service (SNS) users amongst Internet users was 48.8% [5]. 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, with 60.2 million daily active users [6], and has therefore become a valuable source of people's opinions and sentiments.

Zheng Yuan
Computer Laboratory, University of Cambridge, United Kingdom, e-mail:
Zheng.Yuan@cl.cam.ac.uk

Matthew Purver
School of Electronic Engineering and Computer Science, Queen Mary University of London,
United Kingdom, e-mail: m.purver@qmul.ac.uk

Microblog texts (called *statuses* in Weibo) 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-labelled training data. The labelling process is difficult and time-consuming with large datasets, and can be unreliable when attempting to infer an author’s emotional state from short texts [31]. Our solution is to use *distant supervision*: we adapt the approach of [17, 31] to Weibo data, using emoticons and Weibo’s built-in smilies as author-generated emotion labels for training, allowing us to learn a model of the associated language which can 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 Analysis & Emotion Detection

Most research in this area focuses on sentiment analysis – classifying text as positive or negative [27]. 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 [8] and others in using the definition in the work of [11], providing six basic emotions: anger, disgust, fear, happiness, sadness, and surprise.

Algorithms previously used for this task range from matching words in a sentiment lexicon to training classifiers with labelled data. In early work, [41] used mutual information between document phrases and the word “excellent” and “poor” to get the average sentiment orientation of reviews. They used unsupervised classification and achieved an average accuracy of 74%. Phrases containing adjectives or adverbs were extracted and used since they are good indicators of subjective [19]. [28] first applied different machine learning methods to detect the polarity of movie reviews. They reported the effectiveness of using machine learning techniques for sentiment classification: machine learning approach beats human-produced base-

lines easily. However, the performance was not as good as traditional topic-based text classification. They evaluated three machine learning methods (Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVMs)) and results showed that unigram presence information seemed to be the most effective. [45] used movie review comments from social network Digg,¹ and evaluated both supervised learning (NB, ME, Decision trees) and unsupervised learning (K-Means). In addition to a bag-of-words model, they also tried to incorporate WordNet synonyms information. They came to a similar conclusion with [28] that the simple bag-of-words model performs relatively well. [40] proposed a way of using a multiple classifier based on three different classifiers. Results showed that the integrated methods outperformed all three single classifiers.

2.2 Distant Supervision

Distant supervision is an approach which combines standard supervised classification methods with a weakly labelled training dataset; it can be seen as an example of semi-supervised learning in that it exploits large amounts of data without access to expert gold-standard labels. [17, 26], following [32], use emoticons in Twitter messages to provide these weak (or *noisy*) labels, then learn a classifier on the basis of the remaining text (after removal of the emoticons) to classify positive/negative sentiment with above 80% accuracy.

[46] showed that emoticons have an important role in emphasizing the emotions conveyed in a sentence; they can therefore give us direct access to authors' own emotions. [29, 10] similarly found that they tend to increase the intensity of the associated verbal content, rather than replacing it (perhaps playing a similar role to laughter, facial expressions and other non-verbal behaviour). We would therefore expect them to be suitable for use as labels in a distant supervision approach, indexing the emotional content while leaving its verbal expression largely unaffected when the emoticons are removed. [31] investigated the applicability of this approach to English Twitter messages, using a broader set of emoticons to extend the distant supervision approach to six-way emotion classification, and we apply a similar approach here to Chinese Weibo statuses. However, in addition to the widely used, domain-independent emoticons, other markers have emerged for particular interfaces or domains. Weibo provides a built-in set of smilies that can work as special emoticons that help us better understand authors' emotions.

¹ <http://digg.com>

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 [12, 35, 15, 18, 21, 43]. These methods can be roughly classified into two categories: lexicon-based method and character-tagging method.

The idea for lexicon-based method is “segmentation”. The basic technique for identifying distinct words is based on the lexicon-based identification scheme [4]. This approach performs the word segmentation process by 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.

The character-tagging method was first introduced by [44]. It is more like a “word-building” process: it treats the word segmentation as a sequence labeling problem by assigning labels to all characters. Labels indicate whether a character locates at the beginning of, inside or at the end of a word. Several discriminative sequential learning algorithms have been exploited (e.g., conditional random fields (CRFs) [39], latent variable CRFs [37], structured perceptron [20], and the Passive-Aggressive algorithm [36]). However, the performance on social media data is not satisfying as the data is so different from the existing training libraries used.

3 Weibo Corpus

3.1 Corpus Collection

Our training data consisted of Weibo statuses with emoticons or smilies (see Sect. 3.2). Since Weibo has a public API,² training data can be collected through automated means. To use the API, we also need to create a Weibo account and register an application. We wrote a Python script which requested the *statuses_public_timeline API*³ every 30 seconds and inserted the collected data into a *MongoDB*⁴ database. We constructed a corpus of Weibo data, filtering out messages not containing emotion labels (see Sect. 3.2 & 3.4 for details).

² <http://open.weibo.com/wiki/API/en>

³ http://open.weibo.com/wiki/2/statuses/public_timeline/en

⁴ <http://www.mongodb.org/>

3.2 Emotion Labels

Two kinds of emotion labels (emoticons and smilies) were used as noisy labels. By “noisy”, we mean that the emoticons and smilies are noisy themselves compared to gold-standard manual labels: to some degree ambiguous or vague in their meaning. Not all emoticons and smilies are closely related to these six emotion classes considered in our work; and some emoticons or smilies may be used differently in different situations, as people have different understandings. Smilies are Weibo built-in smilies (see Fig. 1) which form a finite, fixed set defined by the Weibo interface. Emoticons here are Eastern-style emoticons, which are made up of several characters and can thus be defined by the user; note that they are very different from Western-style emoticons [23] (see Table 1).

Eastern-style and Western-style emoticons are different, mostly because of different habits from using very different languages. For Western-style emoticons, people are used to reading them from left to right: Western emoticons are generally taken as being rotated by 90 degrees [30]. They are usually made of two to four characters and are of a relatively small number, generally focussing on some feature of mouth shape. Eastern emoticons, in contrast, are usually un-rotated and present faces, gestures, or postures from a point of view easily comprehensible to the reader.

Table 1 Emoticons: Eastern style v.s. Western style

Emotion Classes	Eastern Style	Western Style
Happiness/Smile	(^_^)	:)
Sadness/Cry	(T_T)	:(
Anger	(^-^)	:@

At the beginning, we looked at all Eastern-style emoticons and Weibo built-in smilies available. Initial investigation found that not all emoticons and smilies can be classified into Ekman’s six emotion classes [11]; and for some less frequently used labels, authors have widely different understandings. We therefore identified the most widely used and well-known emoticons/smilies; to then determine whether these would be reliable as labels, we set up a web survey to examine whether people could classify these emoticons/smilies consistently.⁵

Our survey contained two parts. In the first part, we asked people to choose one from the six emotion classes that best matched each of our identified emoticons/smilies. We also provided a *None of the above* option allowing participants to give their own definitions. In the second part, we asked people to tick all the emoticons and smilies they would use to convey each of the six emotions; we also allowed them to fill in other emoticons/smilies of their own that they would use for each emotion class. The survey was distributed via Weibo and only Chinese Weibo

⁵ Available at: <http://www.sojump.com/jq/1935017.aspx?npb=1>













Fig. 1 Screenshot of the first page of Weibo built-in smiles.

users were allowed to take part. 56 individuals completed our survey in two days time and full results are given in Appendix Table 9.

From the results of this, we identified 12 emoticons and 10 smilies to use as emotion labels (see Table 2). It is worth noting that we found no reliable emoticons for *disgust*, nor any reliable labels of either kind for *fear*. One reason may be that both *disgust* and *fear*, as emotion classes, are themselves difficult to represent (as facial expressions) using only punctuation and letters. For *fear*, we even found no relevant smilies in the Weibo interface. We believe this is because there is no obvious distinguishing feature on a *fear* face. In addition, people seem to use other emotions with *fear*, like “nervous”, “cry”. In order to ensure a reliable labelling, we decided to use only one smily for *disgust*, and the keyword 害怕 for *fear* (a Chinese word meaning *fear*). However, we should be careful with keywords as they might not work well. Removing a word from a text may affect the meaning of the message itself and leave the rest of the text less informative and reliable. In addition, words are verbal, so they are subject to things like negation. Using keywords as emotion labels may be less reliable and it may result in lots of false positive examples.

Table 2 Conventional markers used for emotion classes

Emotion Classes	Emoticons	Smilies
Anger	(ㄟ_ㄟ)	 [怒 nù “Anger”]  [怒骂 nù mà “Curse”]
Disgust	N/A	 [吐 tù ”Spit”]
Fear	N/A	N/A
Happiness	(*^_*^*) (*^_*) (*^o^*) o(n_n)o o(^_^)o (^o^) (^-)	 [嘻嘻 xī xī “Hee hee”]  [哈哈 hā hā “Haha”]  [鼓掌 gǔ zhǎng “Applaud”]  [大开心 dà kāi xīn “So happy”]
Sadness	(T.T) (T.T) (π.π)	 [泪 lèi “Tear”]  [悲伤 bēi shāng “Sad”]
Surprise	(OMG)	 [吃惊 chī jīng “Surprise”]

3.3 Text Processing

Initial investigation also found that some Weibo statuses are mixtures of different language units: as well as Chinese, English words were also sometimes present and provided useful information. Therefore, in our work, not only Chinese characters/words, but also any lexical items from other languages were included as features. Weibo usernames (starting with @) and URLs were removed. Punctuation was included as a feature (treated like a lexical unigram), with any repeated punctuations being normalised to 3 characters. We then removed the labelling emoticons and smilies from the texts, using them instead only as 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.

To use word-based features, we need to segment the statuses into words. There are lots of Chinese word segmentation tools; however, many are unsuitable for on-line social media text; we compared *Pymmseg*,⁶ *Smallseg*⁷ and *Stanford Chinese Word Segmenter*,⁸ which all appeared to give reasonable results. *Pymmseg* uses the MMSEG algorithm [38]. *Smallseg* is an open sourced Chinese segmentation tool based on DFA. *Stanford Segmenter* is CRF-based [39].

⁶ <https://code.google.com/p/pymmseg-cpp/>

⁷ <https://code.google.com/p/smallseg/>

⁸ <http://nlp.stanford.edu/software/segmenter.shtml>

3.4 Corpus Analysis

Our corpus contains 1,027,853 Weibo statuses with emotion labels; Table 3 shows statistics. The number of Weibo statuses varied with the popularity of labels themselves: labels for `happiness` and `sadness` are much more frequent than others; very similar results were observed on English Twitter (see e.g., [31]), suggesting that these frequencies are relatively stable across very different languages.

Table 3 Number of Weibo statuses per emotion class

Emotion Classes	Using Emoticons Only	Using Smilies Only	Using Both Labels
Anger	427	60,271	60,698
Disgust	0	8,463	8,463
Fear	Using keyword: 39,978*		
Happiness	19,979	529,077	549,056
Sadness	38,676	307,427	346,103
Surprise	3,097	20,458	23,555

* For “fear”, we used the Chinese keyword 害怕 as the emotion label – see Sect. 3.2.

Overall frequencies show that users of Weibo are more likely to use 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

Machine learning techniques have been shown to be effective for traditional text classification and sentiment analysis. Here, we use Support Vector Machines (SVMs) [42], a state-of-the-art supervised kernel method. The basic idea is to find a maximum-margin hyperplane – a hyperplane that can separate two different classes correctly, and simultaneously maximize the margin (or the distance) between that hyperplane and other “difficult points” close to the hyperplane. These “difficult points” are called support vectors, and the decision function is fully specified by these support vectors. New testing examples are then assigned to one side of the hyperplane. Classifiers trained using SVMs have been shown to have better performance than other classifiers: [22] proved that SVMs consistently achieved good performance on text categorization tasks and outperformed other methods substantially and significantly; [28] applied different machine learning methods to detect the polarity of movie reviews. By evaluating three machine learning methods: Naïve Bayes (NB), Maximum Entropy (ME) and SVMs, they showed that SVMs had the best performance and NB turned out to be the worst. SVMs are good for high-dimensional fea-

ture spaces [22], while, other classifiers are training expensive when dealing with a large number of features.

In our work, classification was using SVMs throughout, with the help of LIBLINEAR [13]. LIBLINEAR inherits many features of LIBSVM [3], but is more efficient for training large-scale problems without using kernels. The performance was evaluated using 10-fold cross validation.

Cross validation is used to estimate how well a model generalises [24]. For one round of cross validation, the dataset is partitioned into two subsets, one for training (*training set*) and one for testing (*validation set* or *testing set*). Several rounds of cross validation are performed, with different partitionings, in order to assess variance. Then we average the results and calculate the standard deviation (σ). F-fold cross validation was introduced by [16]. A single dataset is divided into F chunks; in each fold, 1 chunk is retained as the validation data (*test set*) while the remaining ($F - 1$) chunks are used as training data (*training set*). This process is repeated F times so that each of the F chunks is used exactly once as a test set.

Our training datasets were balanced: a dataset of size N contained $N/2$ positive instances (Weibo statuses containing labels for this emotion class) and $N/2$ negative ones (Weibo statuses containing labels from other classes). For $N/2$ negative instances, we randomly selected instances from other emotion classes for larger datasets ($N > 50,000$), but ensured an even weighting across negative classes for smaller sets to prevent bias towards one negative class.

4.1 Feature Selection

An important part of data-driven approach is converting a piece of text (the “observation”) into a feature vector for text processing. A suitable feature vector should be designed and it should contain as few features as necessary. There is lots of work addressing the feature extraction problem for machine learning (e.g., see [33, 14]). In this section, we focused on two types of lexical features: word-based features and character-based features.

4.1.1 Word-based Features

Chinese is written without spaces between words. In order to identify lexical features, we need to segment them first. Classification performance depends largely on the quality of the lexical features we obtain from different Chinese word segmentation tools.

However, people might find it difficult to apply existing segmentation tools to social media data. On one hand, unconventional words are used in microblogs: misspelt words, cyber words, as well as new words (see e.g. [1]). On the other hand, there are some pre-defined structures which are not used in other domains: Weibo usernames (@username), hashtags (#topic#), URLs, emoticons, smilies, etc.

For these latter unconventional (but known) structures, we can treat them separately, removing them before passing through the segmenter. However, for unconventional and misspelled words, this is not possible in general, and it is difficult for existing tools to identify them correctly. It may require better segmentation algorithms and new models should be trained using social media data. We investigated the effect of three different segmentation tools and results are presented in Fig. 2.

Results showed that *Pymmseg* outperformed *Smallseg* and *Stanford Segmenter* for all emotion classes except *surprise* (where *Stanford Segmenter* yielded the best performance) as training dataset size increased. We can also learn from the results that accuracy increased as we using more training examples (see Sect. 4.2). We also want to point out that in terms of segmentation speed, *Pymmseg* is the fastest and *Stanford Segmenter* is the slowest. Therefore, we used *Pymmseg* for later experiments.

4.1.2 Character-based Features

For character-based features, rather than requiring word segmentation, we simply treat each Chinese character as a unigram feature, as well as each punctuation character, emoticon and smiley (see Table 4).

Table 4 An example of one Weibo status and its n-gram features: repeated punctuations were normalised to 3 chars and reserved as a unigram; smiley was reserved as a unigram; Weibo username was removed. For higher-order n-grams, lower-order n-gram features were also included.

Weibo:	好饿!!!!想吃东西的举手[泪]@飞飞飞鸟_sunshinebird
unigram:	好饿!!!想吃东西的举手[泪]
bigram:	好饿!!!想吃东西的举手[泪] 好饿饿!!!想想吃吃东东西西的的举举手手[泪]
trigram:	好饿!!!想吃东西的举手[泪] 好饿饿!!!想想吃吃东东西西的的举举手手[泪] 好饿!!!饿!!!想想吃吃东吃东西东西的西的举的举手举手[泪]
4-gram:	好饿!!!想吃东西的举手[泪] 好饿饿!!!想想吃吃东东西西的的举举手手[泪] 好饿!!!饿!!!想想吃吃东吃东西东西的西的举的举手举手[泪] 好饿!!!想想饿!!!想想吃吃东吃东西吃东西的东西的举西的举手手的举手[泪]
5-gram:	好饿!!!想吃东西的举手[泪] 好饿饿!!!想想吃吃东东西西的的举举手手[泪] 好饿!!!饿!!!想想吃吃东吃东西东西的西的举的举手举手[泪] 好饿!!!想想饿!!!想想吃吃东吃东西吃东西的东西的举西的举手手的举手[泪] 好饿!!!想想饿!!!想想吃吃东吃东西吃东西的东西的举西的举手手的举手[泪]

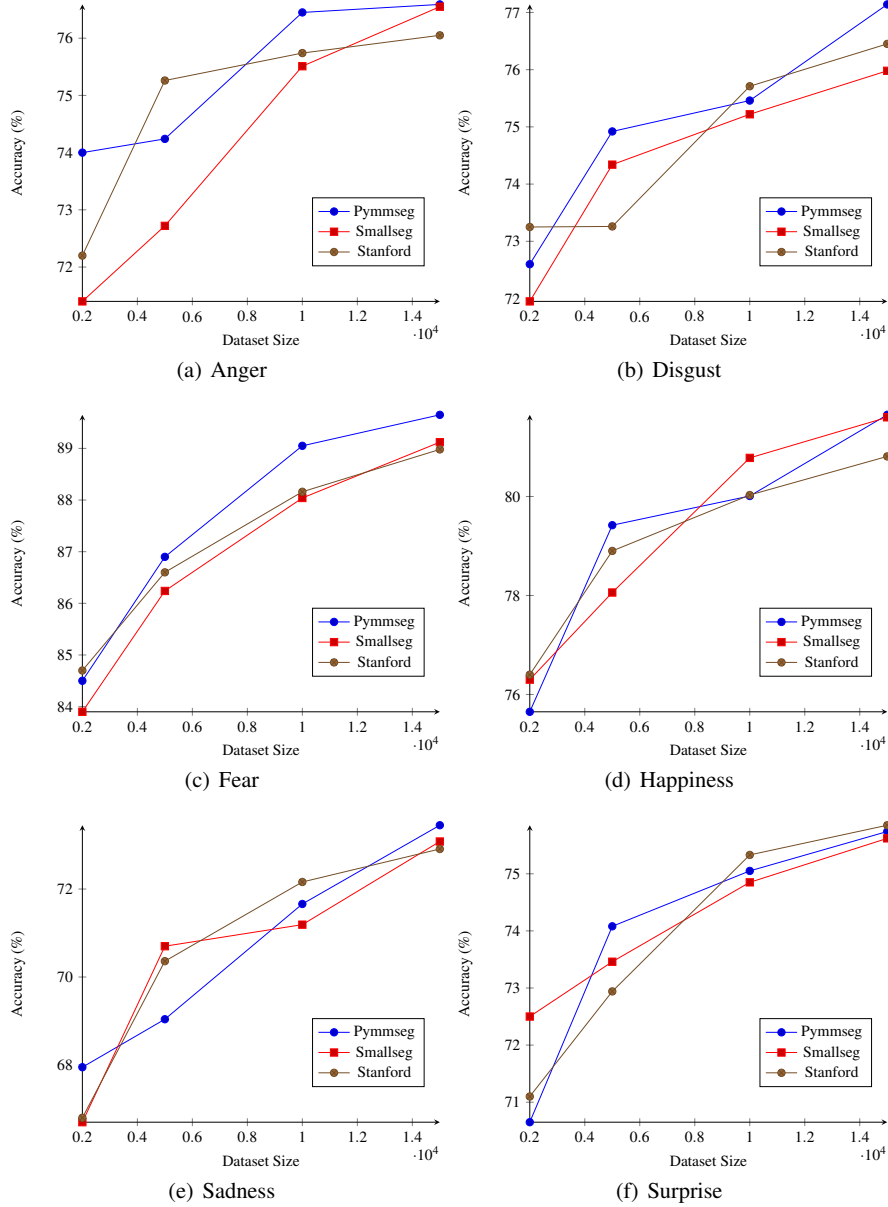


Fig. 2 Classification results of word-based features based on different segmentation tools

Whether higher-order n-grams are useful features appears to be a matter of some debate. [28] reported that unigrams outperformed bigrams when classifying movie

reviews by sentiment polarity, but [9] found that bigrams and trigrams can give better product-review polarity classification.

In our experiments with higher-order n-grams, we also included lower-order n-grams (e.g., for 5-grams, we used all unigrams, bigrams, trigrams, 4-grams and 5-grams as features, see Table 4), as there are lots of Chinese words with only one character.

Results showed that higher-order n-grams are useful features for our wide-topic social media Weibo data. Higher-order n-grams (bigrams, trigrams, 4-grams and 5-grams) outperformed unigrams for all emotion classes by a large margin (see Fig. 3).

We stopped at 5-gram since the accuracy didn't improve any more. And as we adding higher-order n-gram features, it took more time to train classifiers.

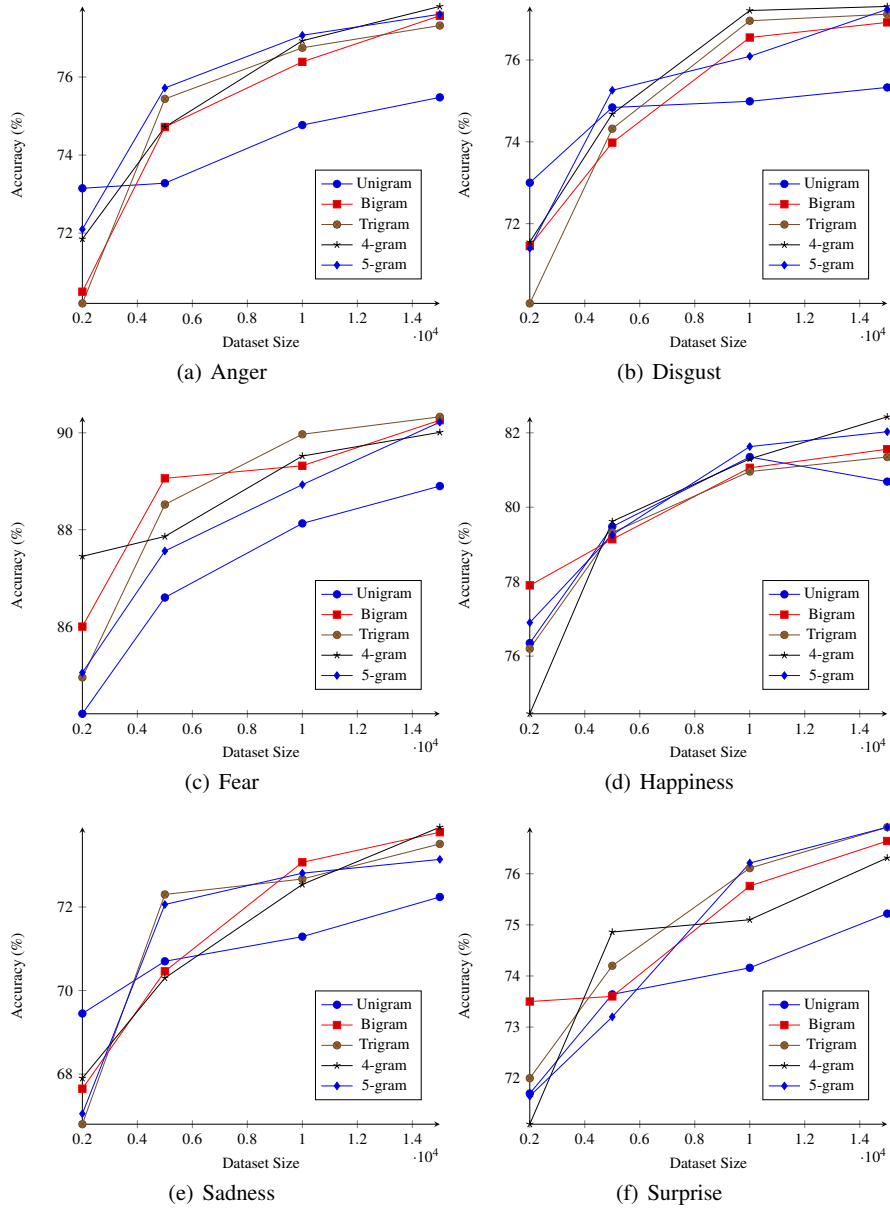


Fig. 3 Classification results of character-based n-gram features

4.1.3 Word-based Features v.s. Character-based Features

Looking at all six emotion classes, we found that word-based features did not beat character-based ones. Character-based higher-order n -gram features had better performance than word-based features (even using the most effective segmenter, *Pymmseg*) for all emotion classes except *sadness* – see Table 5.

Table 5 Classification accuracy for all six emotion classes ($N = 15,000$). The best one for each emotion class is marked in **bold**.

	no. of features*	Accuracy (%)					
		Anger	Disgust	Fear	Happiness	Sadness	Surprise
Word-based (Pymmseg)							
Unigram	45,103	76.59	77.14	89.65	81.65	73.45	75.74
Bigram	161,816	77.31	77.39	89.95	82.04	74.23	76.10
Trigram	261,070	77.21	76.62	90.07	81.79	74.29	76.43
4-gram	331,667	77.01	76.91	90.47	82.20	73.75	76.68
5-gram	394,352	77.49	77.47	90.17	81.97	73.27	76.00
Character-based							
Unigram	12,983	75.90	75.27	88.31	80.53	72.10	75.73
Bigram	139,897	77.17	77.33	90.27	81.77	73.17	75.92
Trigram	339,969	77.06	76.77	90.21	82.23	73.51	77.05
4-gram	498,838	77.29	77.83	90.56	82.36	73.73	75.75
5-gram	616,744	77.89	77.62	90.31	82.08	74.12	76.39

* For higher-order n -grams ($n > 1$), we removed features below a certain frequency threshold ($f = 2$).

Our results suggested that we could just use Chinese characters, rather than doing any word segmentation. Three out of six emotion classes achieved their best performance by using character-based 4-gram features : *disgust*, *fear*, and *happiness*.

Examination of the segmented data showed that these three segmentation tools didn’t work well with our social media data and made lots of segmentation mistakes. In addition, they produced many segmented words which contained only one character. The use of character-based features was therefore preferred and 4-gram features were used in later experiments.

4.2 Increasing Dataset Size

So far, experiments results also showed that increasing dataset sizes increased accuracy up to $N = 15,000$ (see Fig. 2 & 3). In this experiment, we kept increasing training dataset sizes for all six emotion classes and compared their classification results. Character-based 4-gram features were used, and as mentioned before, for larger datasets ($N > 50,000$), we randomly selected negative training examples from other emotion classes (see Sect. 4).

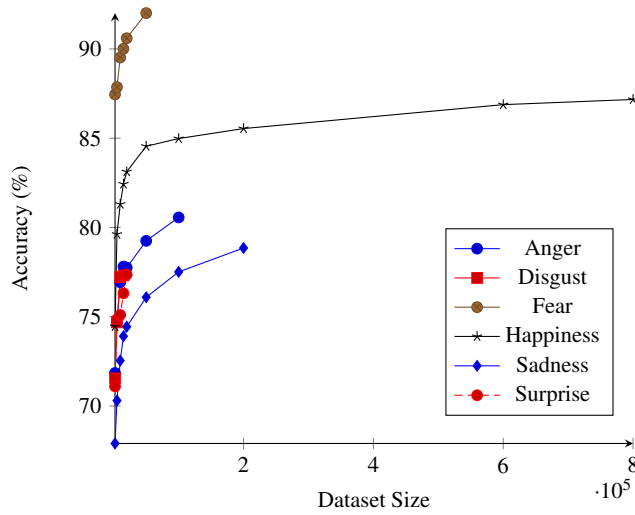


Fig. 4 Classification results for all six emotion classes

Because of the unbalanced number of Weibo statuses for each emotion class (see Sect. 3.4), the largest training dataset size for each emotion class varied: from $N = 15,000$ for *disgust* to $N = 800,000$ for *happiness*. Classification accuracy (using cross-validation) increased as we added more training examples, and does not appear to approach an asymptote until the largest sizes – see Fig. 4 and Table 6. As our dataset sizes increase over time, we therefore expect improvements in accuracy for all six emotion classes.

However, performances are quite different (see Table 6) : *fear* is the most accurately predicted emotion (92.01%) with the keyword as emotion label, followed by *happiness* (87.17%), *anger* (80.56%), *sadness* (78.85%), *surprise* (77.36%) and *disgust* (77.31%).

Table 6 Classification results (accuracy (%)) for all six emotion classes. The best one for each emotion class is marked in bold.

Training Sizes	Anger	Disgust	Fear	Happiness	Sadness	Surprise
2,000	71.85	71.55	87.45	74.45	67.90	71.10
5,000	74.72	74.68	87.86	79.62	70.30	74.86
10,000	76.93	77.21	89.52	81.30	72.54	75.10
15,000	77.81	77.31	90.01	82.43	73.91	76.31
20,000	77.75		90.60	83.12	74.44	77.36
50,000	79.25		92.01	84.55	76.09	
100,000	80.56			84.98	77.51	
200,000				85.54	78.85	
600,000				86.88		
800,000				87.17		

4.3 *Emotion Labels*

In all experiments above, we used a random sample of instances “labelled” with either emoticons or smilies. In this experiment, we compared these two different types of emotion labels (emoticons and smilies) in terms of their classification accuracy. Four kinds of training dataset were constructed and tested for happiness, sadness and surprise:

- A dataset only contained instances collected with emoticons;
- A dataset only contained instances collected with smilies;
- Half of the training examples were collected with emoticons and the other half were collected with smilies;
- The training examples were randomly selected from all the instances collected with both emoticons and smilies.⁹

Comparing the accuracies between these sets tells us which of the label types is used in a more consistent way: association with a more consistent distribution of words/characters will result in higher classification accuracy (accuracy of prediction of emotion label). Results (see Fig. 5) showed that emoticon labels were easier to classify than smilies. By examining a sample of the data directly, we found that people use emoticons in a more systematic or consistent way. They tend to use emoticons to tell others what their real emotions are (happiness, sadness etc.); 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.

4.4 *Manual Labelling*

So far, we used only the distant (“noisy”) labels for both training and testing. In other words, classification accuracy is strictly only a measure of ability to predict the noisy label’s presence (i.e. use of an emoticon or smiley), rather than necessarily measuring the ability to predict the author’s emotion. To examine how well the two correspond, we must test against human judgements.

Amazon’s Mechanical Turk (MTurk)¹⁰ service has shown to be useful for gathering human judgements for many simple NLP tasks (e.g., see [25, 34, 7, 2]). In our final experiment, we used MTurk to collect some manually labelled test data.

Another set of 2,190 instances was used for human annotation. These instances were collected using either emoticons or smilies, and were evenly distributed across our 6 emotion classes. Human annotators were asked to choose the strongest emotion class behind the message, with only one class allowed, although a *None of the*

⁹ That is how we constructed our training datasets for previous experiments.

¹⁰ <https://www.mturk.com/mturk/welcome>

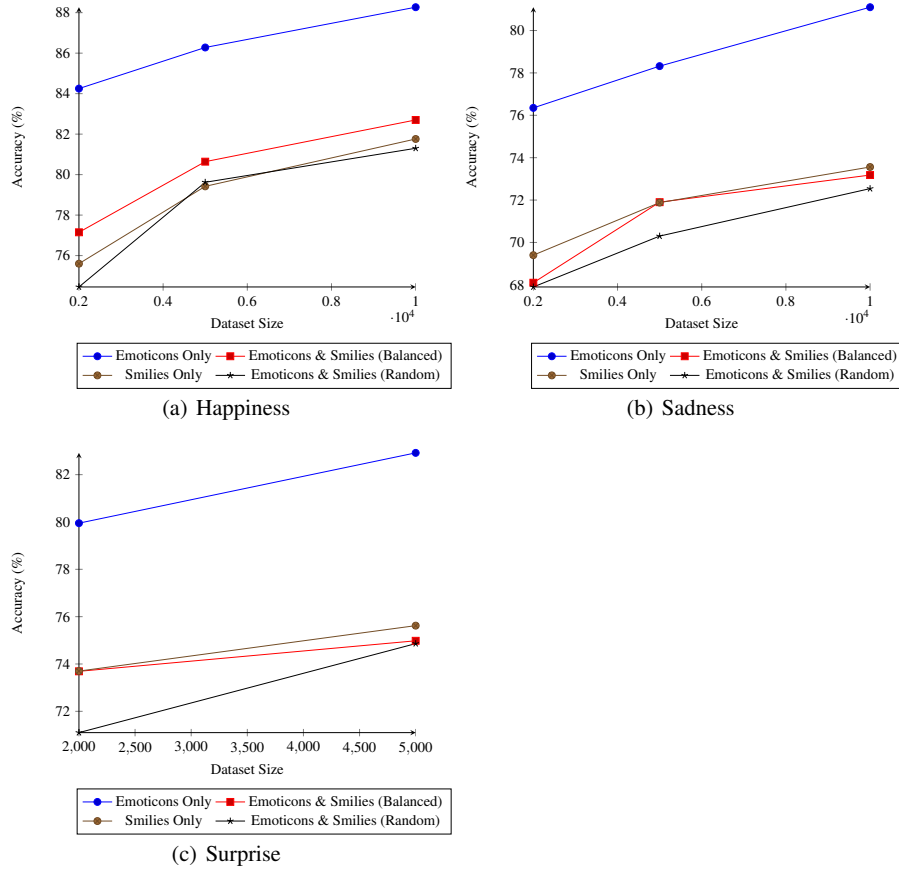


Fig. 5 Comparison of two different types of labels. Character-base 4-gram features were used. Performance was evaluated using 10-fold cross validation.

above option was also provided. Each instance was labelled by three different annotators.

Agreement between annotators was poor: only 26% instances (571 out of 2,190) were assigned the same labels by all three annotators. These unanimous instances were quite unbalanced: from 5 examples for *fear* to 289 examples for *happiness*. When looking at instances agreed by a majority (i.e. at least two annotators), we got 1,335 (out of 2,190) examples varying from 27 for *fear* to 553 for *happiness* – see Table 7.

Two rounds of evaluation were performed where instances agreed by all and majority were used respectively. The best classifier for each emotion class from Sect. 4.2 was used. Since the test dataset was unbalanced, precision, recall and F1 for the class in question were used instead of accuracy. Recall is much higher than precision for some emotions (*sadness*, *surprise*, *disgust* and *fear*) when

Table 7 Number of agreed instances for each emotion class

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	All
test 1 ^a	93	26	5	289	103	55	571
test 2 ^b	216	102	27	553	267	170	1,335

^a labels of instances were agreed by all three annotators

^b labels of instances were agreed by at least two annotators

using default settings. In order to have a consistent F-score to compare between emotion classes, we also tuned these experiments so that recall approximate equals precision. Overall performance is shown in Table 8.

As before, results for `happiness` and `anger` are quite good, which showed that:

1. These two emotion classes are easier to detect;
2. The distant labels used for these two emotion classes are reliable;
3. Our classifiers are able to detect these two emotions.

Results for `surprise`, `sadness` and `disgust` can perhaps be considered reasonable, considering there are far fewer positive examples than negative ones in their test sets.

However, the result for `fear` is poor. Considering the low number of annotated positive test examples (see Table 7), we may conclude that this emotion class is difficult to identify even for human annotators. It is interesting to note that our classifier failed to detect `fear` in these annotated examples even though it achieved high cross-validation accuracy (see Sect. 4.1 & 4.2). This was the only emotion category where we used the presence of a keyword, rather than a non-verbal sign (emoticon or smiley) – this suggests that the use of keywords is a poor method for distant supervision, as suspected.

Table 8 Classification results on manually labelled data

(a) Test on instances agreed by all three annotators

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Precision	90.22%	74.07%	5.26%	94.74%	71.15%	81.82%
Recall	89.25%	76.92%	20.00%	93.43%	71.84%	81.82%
F1	89.73%	75.47%	8.33%	94.08%	71.50%	81.82%

(b) Test on instances agreed by at least two annotators

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Precision	72.43%	61.39%	48.15%	87.70%	64.71%	69.46%
Recall	71.76%	60.78%	48.15%	88.97%	65.92%	68.24%
F1	72.09%	61.08%	48.15%	88.33%	65.31%	68.84%

5 Conclusion

In our work, we used SVMs for automatic emotion detection for Chinese microblog texts. We collected our own Weibo corpus and defined new emoticons and smilies as distant labels. Our results showed that using emoticons and smilies as noisy labels can be an effective way to perform distant supervision for Chinese, while the use of keywords extracted from the text is not effective. Emoticons seem to be more reliable for emotion detection than smilies.

It was also found that, when dealing with social media data, many existing Chinese word segmentation tools do not work well. Instead, we can use characters as lexical features and performance improves with higher-order n-grams. Character-based 4-gram features seem to be the most effective. Increasing the dataset size also improves performance, and our future work will examine larger sets.

Performance for different emotion classes are quite different: *happiness* is the most accurately predicted emotion (87.17%), followed by *anger* (80.56%). The effectiveness of our classifiers for these two emotion classes was also verified by using human annotated test data. Test results on manually labelled data also showed that the other four emotion classes (*sadness*, *surprise*, *disgust* and *fear*) are difficult to classify, either because reliable labels are hard to find (especially in the case of *fear*), and/or because they are difficult to detect even for human annotators.






































Appendix

56 individuals completed our survey; the detailed results are presented here – see Table 9.

References

1. Agichtein E, Castillo C, Donato D, Gionis A, Mishne G (2008) Finding high-quality content in social media. In: Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM'08). pp:183–194
2. Bloodgood M, Callison-Burch C (2010) Bucking the Trend: Large-Scale Cost-Focused Active Learning for Statistical Machine Translation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden. pp:854–864
3. Chang C, Lin C (2001) LIBSVM: a library for Support Vector Machines. Available at: <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf> Cited 4 Feb 2014
4. Chen K, Liu S (1992) Word identification for Mandarin Chinese sentences. In: Proceedings of the 14th conference on Computational linguistics. Vol. 1, pp:101-107
5. China Internet Network Information Center(CINIC) (2013) The 32nd Statistical Report on Internet Development in China. Available at: <http://www1.cnnic.cn/IDR/ReportDownloads/201310/P020131029430558704972.pdf> Cited 2 Feb 2014

Table 9 Survey results showing the percentage of votes each emotion class received for each label. The best match for the defined labels used in our work are marked in **bold**.

Emotion Labels	Anger	Disgust	Fear	Happiness	Sadness	Surprise	None
 [怒 nù "Anger"]	85.71%	1.79%	0	1.79%	0	0	10.71%
 [怒骂 nù mà "Curse"]	73.21%	3.57%	1.79%	1.79%	1.79%	0	17.86%
 [吐 tù "Spit"]	1.79%	58.93%	0	0	1.79%	0	37.50%
 [嘻嘻 xī xī "Hee hee"]	0	0	0	71.43%	0	0	28.57%
 [哈哈 hā hā "Haha"]	0	0	0	80.36%	1.79%	0	17.86%
 [鼓掌 gǔ zhǎng "Applaud"]	0	0	1.79%	73.21%	0	0	25.00%
 [开开心心 kāi xīn "So happy"]	0	0	1.79%	73.21%	0	0	25.00%
 [泪 lèi "Tear"]	1.79%	0	1.79%	0	89.29%	0	7.14%
 [悲伤 bēi shāng "Sad"]	0	1.79%	0	0	89.29%	0	8.93%
 [吃惊 chī jīng "Surprise"]	1.79%	0	3.57%	0	0	76.79%	17.86%
 [哼 hēng "humph"]	50.00%	19.64%	3.57%	1.79%	0	3.57%	21.43%
 [鄙视 bǐ shì "Despise"]	3.57%	35.71%	0	1.79%	0	0	58.93%
 [失望 shī wàng "Disappointed"]	0	1.79%	1.79%	0	53.57%	0	42.86%
 [伤心 shāng xīn "Sad"]	78.57%	7.14%	0	0	1.79%	0	12.50%
 [失望 shī wàng "Disappointed"]	39.29%	23.21%	0	0	14.29%	0	23.21%
 [伤心 shāng xīn "Sad"]	1.79%	14.29%	10.71%	3.57%	33.93%	3.57%	32.14%
 [伤心 shāng xīn "Sad"]	16.07%	3.57%	1.79%	3.57%	23.21%	0	51.79%
 [伤心 shāng xīn "Sad"]	3.57%	0	0	92.86%	0	0	3.57%
 [伤心 shāng xīn "Sad"]	1.79%	0	1.79%	85.71%	0	0	10.71%
 [伤心 shāng xīn "Sad"]	0	0	1.79%	87.50%	0	1.79%	8.93%
 [伤心 shāng xīn "Sad"]	0	0	0	89.29%	0	0	10.71%
 [伤心 shāng xīn "Sad"]	3.57%	0	0	87.50%	0	0	8.93%
 [伤心 shāng xīn "Sad"]	1.79%	1.79%	0	87.50%	1.79%	3.57%	3.57%
 [伤心 shāng xīn "Sad"]	1.79%	1.79%	1.79%	89.29%	1.79%	0	3.57%
 [伤心 shāng xīn "Sad"]	3.57%	0	1.79%	14.29%	1.79%	3.57%	8.93%
 [伤心 shāng xīn "Sad"]	7.14%	3.57%	0	3.57%	60.71%	7.14%	17.86%
 [伤心 shāng xīn "Sad"]	14.29%	0	3.57%	12.50%	57.14%	0	12.50%
 [伤心 shāng xīn "Sad"]	1.79%	3.57%	0	3.57%	80.36%	1.79%	8.93%
 [伤心 shāng xīn "Sad"]	1.79%	0	0	1.79%	1.79%	89.29%	5.36%
 [伤心 shāng xīn "Sad"]	0	1.79%	5.36%	10.71%	0	53.57%	28.57%
 [伤心 shāng xīn "Sad"]	3.57%	0	1.79%	0	1.79%	39.29%	53.57%
 [伤心 shāng xīn "Sad"]	3.57%	3.57%	0	1.79%	1.79%	57.14%	32.14%
 [伤心 shāng xīn "Sad"]	1.79%	1.79%	3.57%	1.79%	1.79%	55.36%	33.93%
 [伤心 shāng xīn "Sad"]	0	0	3.57%	12.50%	1.79%	28.57%	53.57%
 [伤心 shāng xīn "Sad"]	3.57%	1.79%	5.36%	16.07%	5.36%	14.29%	53.57%
 [伤心 shāng xīn "Sad"]	19.64%	3.57%	7.14%	5.36%	8.93%	25.00%	30.36%
 [伤心 shāng xīn "Sad"]	3.57%	8.93%	3.57%	5.36%	26.79%	16.07%	35.71%

6. China, SINA Corporation (SINA) Q3 2013 Earnings Conference Call (2013) Available at: <http://seekingalpha.com/article/1835112-sina-corporations-ceo-discusses-q3-2013-results-earnings-call-transcript> Cited 2 Feb 2014
7. Callison-Burch C (2009) Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazons Mechanical Turk. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009). Singapore. pp:286-295
8. Chuang Z, Wu C (2004) Multimodal emotion recognition from speech and text. In: Computational Linguistics and Chinese Language, 9(2):45–62
9. Dave K, Lawrence S, and Pennock D M (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: WWW2003. pp:519–528
10. Derks D, Bos A, von Grumbkow J (2008) Emoticons and Online Message Interpretation. *Social Science Computer Review* 26(3):379–388
11. Ekman P (1970) Universal facial expressions of emotion. In: *California Mental Health Research Digest*, Vol. 8, Number 4
12. Fan C, Tsai W (1988) Automatic word identification in Chinese sentences by the relaxation technique. In: *Computer Processing of Chinese and Oriental Languages*
13. Fan R, Chang K, Hsieh C, Wang X, and Lin C (2008) LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research* 9(2008), 1871-1874.
14. Forman G (2003) An extensive empirical study of feature selection metrics for text classification. In: *Journal of Machine Learning Research*, 3:1289–1305
15. Gan K, Palmer M, and Lua K (1996) A statistically emergent approach for language processing: Application to modeling context effects in ambiguous Chinese word boundary perception. In: *Computational Linguistics*, 22(4):531-53
16. Geisser S (1975) The predictive sample reuse method with applications. In: *Journal of the American Statistical Association*. pp:320-328
17. Go A, Bhayani R, and Huang L (2009) Twitter Sentiment Classification using Distant Supervision. Master's thesis, Stanford University
18. Guo J (1997) Critical tokenization and its properties. In: *Computational Linguistics*, 23(4):569-596
19. Hatzivassiloglou V, Wiebe J M (2000) Effects of adjective orientation and gradability on sentence subjectivity. In: Proceedings of the 18th International Conference on Computational Linguistics.
20. Jiang W, Huang L, and Liu Q (2009) Automatic adaptation of annotation standards: Chinese word segmentation and pos tagging a case study. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore. pp:522–530
21. Jin W, Chen L (1998) Identifying unknown words in Chinese corpora. In: First Workshop on Chinese Language, University of Pennsylvania, Philadelphia
22. Joachims T (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proceedings of the 10th European Conference on Machine Learning (ECML'08). pp:137–142
23. Kayan S, Fussell S R, and Setlock L D (2006) Cultural differences in the use of instant messaging in Asia and North America. In: Proceedings of the 20th anniversary conference on Computer supported cooperative work (CSCW'06). Banff, Alberta, Canada. pp:525–528
24. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). Morgan Kaufmann, San Mateo, CA
25. Nakov P (2008) Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. In: Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems and Applications (AIMSA 2008). pp:103–117
26. Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta
27. Pang B, Lee L (2008) Opinion mining and sentiment analysis. In: *Foundations and Trends in Information Retrieval*

28. Pang B, Lee L, and Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of Empirical Methods in Natural Language Processing*. pp:79–86
29. Provine R, Spencer R, Mandell, D (2007) Emotional Expression Online: Emoticons Punctuate Website Text Messages. *Journal of Language and Social Psychology*, 26(3):299–307
30. Ptaszynski M, Maciejewski J, Dybala P, Rzepka R and Araki K (2010) CAO: A Fully Automatic Emoticon Analysis System Based on Theory of Kinesics. In: *Affective Computing, IEEE Transactions*
31. Purver M, Battersby S (2012) Experimenting with Distant Supervision for Emotion Classification. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Avignon, France. pp:482–491
32. Read J (2005) Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: *Proceedings of the ACL Student Research Workshop*. Ann Arbor, Michigan. pp:43–48
33. Sebastiani F (2002) Machine learning in automated text categorization. In: *ACM Computing Surveys*, 34(1):1–47
34. Snow R, O'Connor B, Jurafsky D, and Ng A Y (2008) Cheap and Fast But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*. Honolulu, Hawaii.
35. Sproat R, Shih C (1990) A Statistical Method for Finding Word Boundaries in Chinese Text. In: *Computer Processing of Chinese and Oriental Languages*
36. Sun W (2010) Word-based and characterbased word segmentation models: Comparison and combination. In: *Coling 2010: Posters*. Beijing, China. pp:12111219
37. Sun X, Zhang Y, Matsuzaki T, Tsuruoka Y, and Tsujii J (2009) A discriminative latent variable Chinese segmenter with hybrid word/character information. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado. pp:56–64
38. Tsai C (2000) MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm. Available at: <http://technology.chtsai.org/mmseg/> Cited 4 Feb 2014
39. Tseng H, Chang P, Andrew G, Jurafsky D, and Manning C (2005) A Conditional Random Field Word Segmenter. In: *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*
40. Tsutsumi K, Shimada K, and Endo T (2007) Movie Review Classification Based on a Multiple Classifier. In: *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC)*
41. Turney P D (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia. pp:417–424.
42. Vapnik V N (1995) *The Nature of Statistical Learning Theory*
43. Wu A (2003) Customizable segmentation of morphologically derived Words in Chinese. In: *Computational Linguistics and Chinese Language*
44. Xue N (2003) Chinese word segmentation as character tagging. In: *International Journal of Computational Linguistics and Chinese Language Processing*
45. Yessenov K, Misailovic S (2009) Sentiment Analysis of Movie Review Comments. In: *Methodology (2009)*:1-17
46. Yuasa M, Saito K, and Mukawa N (2006) Emoticons convey emotions without cognition of faces: an fMRI study. In: *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. pp:1565–1570