Lexicon-based Sentiment Analysis on Topical Chinese Microblog Messages

CUI Anqi†, ZHANG Haochen, LIU Yiqun, ZHANG Min, MA Shaoping

State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084
† Corresponding author, E-mail: cuianqi@gmail.com

Abstract Microblogging is a popular social media where people express their opinions and sentiment on social topics. The Chinese microblogging service, called *Weibo*, has become a remarkable media in the Chinese society. People are eager to know others' attitudes towards social events, thus sentiment analysis on those topical microblog messages is important. In this paper we introduce a lexicon-based sentiment analysis method. We construct a *Weibo Lexicon* with representative topical words and out-of-vocabulary (OOV) words, which are usually informal and are not existing in formal dictionaries. In addition, we use a propagation algorithm to automatically assign sentiment polarity scores to the words discovered. These scores are more closely reflecting the Weibo context since words may have new or opposite polarities instead or their formal meanings. Evaluations on the classification tasks show that our method is effective on recognizing the subjectivity and sentiment of Weibo sentences. The Weibo lexicon increases the performance of the classifications.

Key words sentiment analysis; subjectivity; microblog; Weibo; sentiment lexicon; out-of-vocabulary (OOV) words

基于微博情感词典的中文微博客话题情感分析

崔安颀† 张昊辰 刘奕群 张敏 马少平

智能技术与系统国家重点实验室，清华信息科学与技术国家实验室（筹），
清华大学计算机科学与技术系，北京 100084；† 通信作者，E-mail: cuianqi@gmail.com

摘要 提出基于微博情感词典的中文微博客话题情感分析方法。使用有代表性的话题相关词语与未登录词，构建微博情感词典，可涵盖传统正式词典中缺失的非正式用语。利用标签迭代技术，自动计算微博情感词典中的词语情感得分。在微博客的上下文环境中，词语可能具有新含义或与传统含义相反。通过自动计算得到的情感得分更接近微博客中体现出的情感倾向。分类任务的评价结果表明使用微博情感词典可帮助识别微博句子的主观性和情感倾向。

关键词 情感分析；主观性；微博客；微博；情感词典；未登录词

Microblogging is a popular User-Generated Content (UGC) service in the Web 2.0 era. Different from a traditional web article, a microblog message has a limited length of content, typically 140 characters. This results in a faster composition, thus users participate in microblogging more often. Microblogging has become a significant media and a rich corpus of users' emotions and opinions, especially on hot topics and events. Since then, researchers have paid much attention to microblog messages.

The textual content of a microblog message is different from traditional web texts, mainly because of the length limit. For example, the topic in a shorter piece of text is usually more focused, and the expressed emotion (or opinion) is more straight forward. This makes it easier for lexicon-based analysis. However, words and expressions used in microblog messages are less formal. They contain abbreviations and out-of-vocabulary (OOV) words which make it more difficult to understand the content. Thus, microblog messages as informal short texts

国家 863 高科技项目（2011AA01A207），自然科学基金（60903107, 61073071），高等学校博士学科点专项科研基金（20090002120005）资助
本文工作完成于清华大学—新加坡国立大学下一代搜索技术研究中心（NExT Search Centre），该中心受到新加坡国立研究基金会（Singapore National Research Foundation）与新加坡媒体发展局交互数字媒体研发项目办公室（Interactive Digital Media R&D Program Office, MDA）联合资助（WBS: R-252-300-001-490）

have become a new interest to the researchers.

Currently the most popular microblogging service in the world is Twitter (http://twitter.com/) which has more than 500 million users as of April 2012[1]. The Twitter messages (called *tweets*) are mostly in English, hence most studies focus on English tweets analysis. Unfortunately, Twitter is not accessible in mainland China. As an alternative, Chinese microblogging services (called *Weibo*) have attracted the Internet users in China. The most popular Weibo services include Sina Weibo (http://weibo.com/), Tencent Weibo (http://t.qq.com/), etc. As of June 2012, 274 million (50.9%) Internet users in China have used Weibo services[1], within only three years since the sites' opening to the public. Though Weibo is new in China's Internet services, it has now become an influential media to the society. Many companies, celebrities and governments are eager to know what people are talking about in Weibo, especially people's opinion against some specific topic. To the opposite, research on Weibo analysis is still preliminary. Besides the fact that Weibo gets popular much later than Twitter, some characteristics of Chinese microblogs different from English microblogs make it more difficult on textual analysis:

1. Chinese as a character-based language contains more information than English with a same length. In Weibo, each message is limited to 140 Chinese characters, but it contains more words and sentences than an English message with 140 alphabets. Longer contents lead to the fact that sentences in one message may or may not be coherent; sometimes they express different topics or even different opinions. Fig. 1 shows an example Weibo message on the topic of "Teachers in China get almost the lowest salary in the world". The message has four sentences with a total of 71 Chinese characters, while its English translation has more than 240 characters. The sentences have different sentiment polarities, implying that a finer granularity (sentence level) is necessary.

S#	Original Text	English Translation	P
1	#中国教师收入全球几垫底#没有几个。	#Teachers in China get almost the lowest salary in the world# Not so many.	φ
2	妈妈今天还说当老师很不容易。	Today mom said it isn't easy to be a teacher.	+
3	不过老师虽然累，但有一群她的学生们。	A teacher works hard and is tired, but she has her students.	-
4	好老师的心态永远、每天都会是：累并快乐着！	A good teacher should always keep tired while feeling happy!	+

Fig. 1 An example Weibo message with four sentences (S#: Sentence number. P: Sentiment polarity, φ for neutral, + for positive and – for negative). Different sentences have different sentiment polarities.

2. Chinese word segmentation is a big challenge in Chinese text analysis. Existing Chinese word segmentation tools and part-of-speech (POS) taggers usually work well on formal texts. However, Internet texts are often informal with many OOV words. Moreover, formal words in Internet contexts may have opposite meanings against their original meanings. Existing algorithms usually fail under these situations. For sentiment analysis, current sentiment lexicons contain only the formal words and their formal polarities. In Internet texts, we need to extend the existing lexicons to catch more sentiment words and expressions, and discover their "new" meanings. Fig. 2 shows two examples of a topic "Crazy green onions" where people complain about the high prices of green onions.

M#	Original Text	English Translation
1	#疯狂的大葱#大葱它肿么了？	#Crazy green onions# What happened to the green onions?
2	#疯狂的大葱#什么都涨，就工资不涨。等工资涨了，什么就都又涨了。鸭梨，，，	#Crazy green onions# Every price is going up except salary. When salary increases, all others are rising again. Pressure…

Fig. 2 Example Weibo messages with new words or meanings (M#: Message number. New words and their translations are underlined). The OOV word "肿么了" pronounces similar as "怎么了" (what happens to). The word "鸭梨" (pear) sounds like "压力" (pressure) so it is often used in Weibo as self-deprecation.

[1] http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842

Facing these problems, on one hand, in this paper we restrict our Weibo sentiment analysis problem to the sentence level analysis on topical Weibo messages. This treatment removes the diverse irrelevant topics in Weibo, and is more useful in practice. On the other hand, our methodology specially discover words from the Weibo corpus. These words are then assigned Weibo-based polarity scores, which represent the sentiment polarities in the context of Weibo instead of formal texts.

The paper is organized as follows: Section 1 gives a formal definition of our problems. Section 2 and 3 introduce the algorithm of constructing the Weibo lexicon and classification with them. Section 4 shows the experiments and evaluations. Section 5 introduces related work. Finally, Section 6 concludes the paper and raises possible future work.

1 Problem Definition

In this paper, we conduct a two-stage sentiment analysis on topical Chinese Weibo messages. The messages are collected with hashtags (topical words or phrases quoted by a pair of # symbol). These messages contain less spam and are relevant to the specified topic. The messages are then divided into sentences for analysis.

Task 1: Given a sentence in a Weibo message, we find out whether or not it is an *opinionated* sentence. Opinionated sentences are the ones that have comments (opinions) on some specific objects, exclusive of personal feelings, wishes and moods. For example, “I am happy” is not opinionated; “I love iPhone’s screen effects” is opinionated.

Task 2: Given an opinionated sentence, we find out its sentiment polarity, i.e., if the opinion is positive, negative or others. Fig. 1 shows some examples of the polarities of Weibo messages.

Note Task 1 is served as the first stage of Task 2. Our evaluations of both tasks consider the ground truth of opinionated sentences.

2 Lexicon Construction

As mentioned before, Chinese Weibo messages have many OOV words and words that have new meanings. Thus we deploy algorithms to automatically discover them and their sentiment polarities, which form the Weibo lexicon.

We first pick out some words as the entries of the lexicon. These entries come from the mid-frequency words and OOV words in our corpus. Then we use a label propagation algorithm to assign polarity scores to them.

2.1 Representative Topical Mid-frequency Words

We collect a background corpus from Tencent Weibo. A total of 849,783 Weibo messages (of 2,264,464 sentences) are collected, covering three months prior to the 22 topics in our training and test datasets. The words in these messages are not concentrated on those topics.

We also collect an extended set of the 22 topics in Tencent Weibo. A total of 44,603 Weibo messages (of 108,113 sentences) are collected. The distribution of words in these messages are much closer to the one in the training and test datasets.

After the messages are cut into sentences, word segmentation (with the ICTCLAS tool[2]) is applied to generate a preliminary segmentation result, together with their POS tags. To overcome the incorrect segmentation, we also generate bi-words and tri-words in addition to the uni-words.

The top-50 *n*-gram words in each set (uni-words, bi-words and tri-words) from the background corpus are considered as high-frequency words, mainly stop words. The *n*-gram words with frequency lower than three in the extended topic corpus are considered as low-frequency words, mainly username or proper nouns. From the extended set, we remove the high- and low-frequency words and get the mid-frequency words, which represent the corresponding topic.

2.2 OOV Words

The OOV words are discovered with context entropy gain and mutual information[3]:

$$Entropy(W) = -\sum_{i \in Next(W)} p_i \ln p_i$$

where $Next(W)$ is the adjacent word (character) of the original word W and p_i is the probability of that word.

These OOV words are grown from characters in the extended topic corpus, hence they are independent to the segmentation results which are not so reliable in Weibo texts. The OOV words serve as a compliment to the word segmentation results.

Note the OOV words may contain a shorter word that is a substring of another longer word. When generating the features, we match the longer words first. If a longer word matches in the sentence, all of its substrings are not considered.

2.3 Sentiment Polarity Assignment

We construct a co-occurrence graph to propagate polarity scores to the words in the lexicon. The mid-frequency words and the OOV words are nodes while their co-occurrences in the Weibo sentences are edges. The weights of the edges are the numbers of co-occurrences between the two nodes.

A public sentiment dictionary[4], containing 728 positive words and 933 negative words (all are formal words), is used for seeds for label propagation. The propagations are conducted twice, one for positive scores (starts with the positive seeds) and one for negative scores (starts from the negative seeds). The label propagation step is similar to [5], which assigns a score to each word by:

$$x^{(n+1)} = x^{(n)} \times W \div (\sum \sum W_{ij})$$

where x is the node vector and W is the adjacent matrix of the co-occurrence graph. Starting with the vector x_0 where only the seeds has a score of one, all the rest nodes are propagated with scores after iterations. The score of seeds are reset to one again between two iterations. Table 1 lists some example words (exclusive of the seeds) and their scores.

Table 1 Example Non-seed Words with Polarity Scores

Word	Translation	Positive Score	Negative Score
蛊惑人心	Demagogy	0.521	0.759
嚣张拔扈	Typo of 嚣张“跋”扈	0.493	0.728
嚣张跋扈	Arrogant and domineering	0.572	0.759
宋祖德	A celebrity who always criticize others in public thus has a low reputation	0.584	0.757
/心碎	Icon of a breaking heart	0.544	0.728
性感	Sexy	0.792	0.626
不拘一格	Not restricted to rigid rules	0.757	0.588
李娜	A top-10 tennis player, the first from China to win a Grand Slam in singles	0.744	0.579
/给力	Icon of 给力, a new word in Internet meaning “awesome”	0.763	0.627

From the table we see that the algorithm automatically recognizes typos, new words and assign polarity scores to them as well as celebrities. Positive words have a higher positive score than negative score, and vice versa. Typos and new words are common in Internet texts, especially in Weibo, thus this methodology is helpful for Network Informal Language analysis.

The polarity scores are within the range of [0; 1]. Words with scores higher than a given threshold are considered as positive (or negative) words. In our work we use multiple thresholds to generate multiple lexicons. For example, a threshold of 0.5 excludes “嚣张拔扈” from the negative lexicon while a lower threshold keeps it.

3 Subjectivity and Sentiment Classification

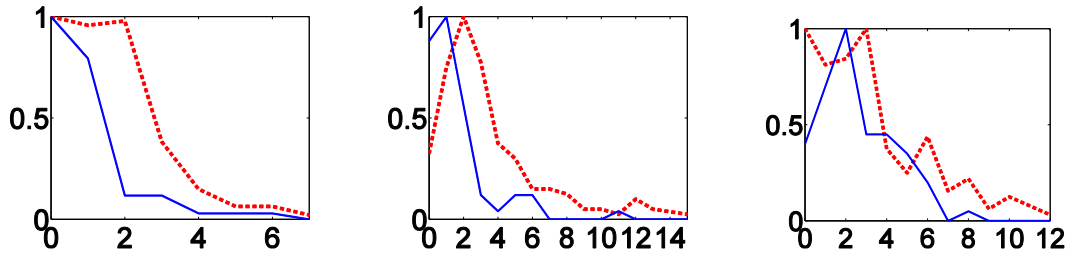
In this section we introduce our supervised classification method.

3.1 Feature Candidates

Three categories of features are chosen as candidate features for the classifier:

1. Non-semantic features: Length (number of characters) of the sentence. Word count (by segmentation).
2. POS features: Word count of each POS tag. Some POS tags are more predictive to the subjectivity such as pronouns, nouns and verbs, as shown in Fig. 3. For example, sentences with no noun are more likely to be opinionated compared to sentences with two nouns.

   One set of the POS features are directly from the ICTCLAS output, i.e., all words are counted into the POS frequencies. The other set is from the Weibo lexicon. In this set we only consider the words that occur in this lexicon. Their POS (or bi-words and tri-words) frequencies are contributed to the features. In this way, the high-(and low-) frequency words do not influence the POS count.



(a) Pronoun (b) Noun (c) Verb

Fig. 3 Normalized number of opinionated (dashed red) sentences and non-opinionated (blue) sentences to the number of words of a specified POS in each sentence. Distributions are computed from the training set.

3. Sentiment dictionaries: We collect three public sentiment dictionaries from Hownet[1], a Student Dictionary[4], and National Taiwan University Sentiment Dictionary[6] (NTUSD). Table 2 lists the statistics of these dictionaries. Features are the number of positive (and negative) words appeared in the dictionaries. Hence $3 \times 2 = 6$ features are generated.

   In addition, we use the Weibo lexicon to generate more sentiment features. As mentioned before, different thresholds generate different word lists, hence the word counts are different. We use them as individual features.

Table 2 Number of entries of the Sentiment Dictionaries

Dictionaries	Positive Count	Negative Count
Hownet	836	1,254
Student Dictionary	728	933
NTUSD	2,810	8,276

A feature selection algorithm is applied to remove the redundant features from the candidates.

3.2 Classification Method

A provided training set is used to train the classifier. In this work we use libSVM[7] with its default settings as the classifier. We first classify the subjectivity of each sentence (Task 1). The subjective (opinionated) sentences are used for sentiment classification (Task 2). A challenge is that most topics are social events where

[1] http://www.keenage.com/html/c_bulletin_2007.htm

people tend to criticize on them. The training set of Task 2 is imbalanced: One commercial topic has 41 positive and 59 negative sentences (of all the opinionated sentences), while the other social topic have negative sentences only (170 sentences). We combine the two topics together as a full training set to decrease the imbalance.

4 Experiments and Evaluations

4.1 Corpus Details

The corpus is provided from *Tencent Weibo* by the Weibo Sentiment Analysis Evaluation Tasks of the First Conference on Natural Language Processing & Chinese Computing (NLP&CC 2012). The training set contains two topics, with a total of 205 Weibo messages with 464 sentences. The test set contains 20 topics, with 17,518 messages of 31,675 sentences. Among them, 1,908 sentences in 10 topics have been annotated. The labels are annotated by experts and are considered as gold standards. Table 3 lists the proportions of classes in each topic.

Table 3 Proportions of Classes in Topics

Dataset	Topic #	Opinionated		Sentiment Polarity		
		Yes	Total	Pos	Neg	Others
Training	1	170	242	0	170	0
	2	101	222	41	59	1
	Total	271	464	41	229	1
Test	1	126	220	5	121	0
	2	128	176	21	106	1
	3	132	222	21	111	0
	4	115	147	8	107	0
	5	123	135	110	13	0
	6	133	217	25	105	3
	7	112	147	2	110	0
	8	122	201	8	110	4
	9	142	230	10	129	3
	10	135	213	24	109	2
	Total	1,268	1,908	234	1,021	13

4.2 Feature Selection

Features that do not bring information gain in the training set are removed to improve the performance.

$$\mathrm{InfoGain}(C, F) = \mathrm{Entropy}(C) - \mathrm{Entropy}(C|F)$$

where *C* is the class and *F* is the feature (attribute).

The selected features are:

1. Non-semantic features: Length of the sentence. Word count.
2. POS features: Adjectives, attributive words, adverbs, numerals, nouns, quantifiers (measure words), auxiliary words and verbs from both the ICTCLAS POS tags and the Weibo lexicon. ICTCLAS's pronouns and punctuations, Weibo lexicon's bigrams and trigrams are also included.
3. Sentiment dictionaries: Negative word count of all the dictionaries, both public resources and our sentiment lexicons. The negative words are important here due to the imbalanced training set.

4.3 Classification Results

Within the test set of more than 30 thousand sentences, experts have annotated 1,761 labels. These gold standards are used to evaluate our performance.

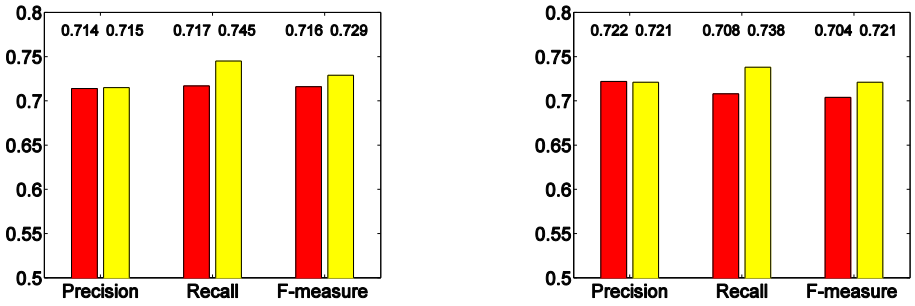
Evaluation measurements are precision, recall and F-measure:

$$\text{Precision} = \frac{\#system\_correct(opinion = Y)}{\#system\_proposed(opinion = Y)}$$

$$\text{Recall} = \frac{\#system\_correct(opinion = Y)}{\#gold(opinion = Y)}$$

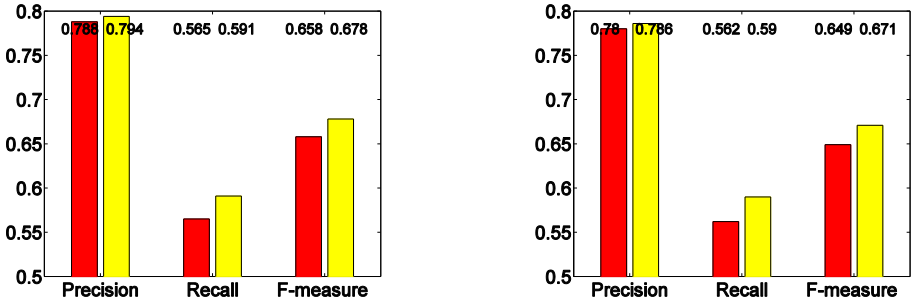
$$\text{F} - \text{measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

We compute the micro-average (performance on the whole dataset) and macro-average (average of performances on each topic). Moreover, we compare the results with and without the features of our Weibo lexicons. Fig. 4 and Fig. 5 illustrates the classification performances on subjectivity and sentiment polarity. The figures show that adding our Weibo lexicon features increases the classification performances.



(a) Micro-average (b) Macro-average

Fig. 4 Comparison of without (left red bar) and with (right yellow bar) the Weibo lexicon features on subjectivity classification.



(a) Micro-average (b) Macro-average

Fig. 5 Comparison of without (left red bar) and with (right yellow bar) the Weibo lexicon features on sentiment classification.

The method on subjectivity classification produces similar precision and recall rates. For sentiment classification, however, precisions are higher than recalls. The reason is similar as the one causing errors in the subjectivity task, where the mis-classified sentences are usually with positive sentiment. Due to the imbalanced training set, positive information is not learned correctly, thus these sentences are often recognized as non-opinionated.

5 Related Work

Sentiment analysis is important in Web text analysis. Traditional methods include building statistical models with machine learning techniques[8][9]. Part-of-speech tags are also used for rule-based approaches[10]. These traditional studies usually focus on formal or longer texts (such as product reviews), where words are formal and their POS tags are reliable.

In Twitter sentiment analysis, researchers have been using some specific features such as emoticons[11] or irregular spellings[5]. However, Chinese is not a spelling language. People do not use English emoticons very often,

since they need to switch their input method from typing Chinese characters to English alphabets. In addition, irregular spellings (such as repeating letters) are less common. These characteristics restrict us from these studies.

In Chinese microblogs, Weibo, instead people are using emotional icons provided by the Weibo websites. People choose icons on the web interface to insert them into their posting messages. Thus, these icons are clues for sentiment analysis[12]. However the icons are not widely used in most Weibo messages; we still need to find a way to analyze the text-only messages.

6 Conclusions and Future Work

In this paper, we propose an algorithm for sentiment analysis on topical Chinese microblog (Weibo) messages. The method is based on the Weibo lexicon, which is automatically generated from the microblog corpus. It contains many new words and OOV words which help recognize the sentiment of informal texts.

Weibo messages have many typos and new words. Formal words may also have different meanings from their original meanings. Therefore, we build a lexicon from a background Weibo corpus and a topical Weibo corpus, to collect potential sentiment words or n-grams that reflect the topics. A label propagation algorithm is applied to assign sentiment polarity scores to the words we have discovered. In this way, the actual polarities of the words with respect to the Weibo context are assigned.

Topical Weibo messages usually have imbalanced sentiment polarities. Many social events lead to a huge number of negative opinions. Therefore, training data is highly imbalanced. This also limits our supervised algorithm. In future work, we will try to find some balanced corpus to construct the lexicon. Some other models for imbalanced training, such as one-class SVM, can also be considered. Moreover, we would like to evaluate the lexicon itself, for how precise the scores are assigned to the new words.

References

[1] CNNIC: The 30th china internet development report. Tech. rep., China Internet Information Center (2012)

[2] Zhang, H.P., Yu, H.K., Xiong, D.Y., Liu, Q.: Hhmm-based chinese lexical analyzer ictclas. In: Proceedings of the second SIGHAN workshop on Chinese language processing - Volume 17. pp. 184—187. SIGHAN '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003)

[3] Li, Z., Zhang, M., Ma, S., Zhou, B., Sun, Y.: Automatic extraction for product feature words from comments on the web. In: Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology. pp. 112—123. AIRS '09, Springer-Verlag, Berlin, Heidelberg (2009)

[4] Zhang, W., Liu, J., Guo, X.: Positive and Negative Words Dictionary for Students. Encyclopedia of China Publishing House (2004)

[5] Cui, A., Zhang, M., Liu, Y., Ma, S.: Emotion tokens: bridging the gap among multilingual twitter sentiment analysis. In: Proceedings of the 7th Asia conference on Information Retrieval Technology. pp. 238—249. AIRS'11, Springer-Verlag, Berlin, Heidelberg (2011)

[6] Ku, L.W., Chen, H.H.: Mining opinions from the web: Beyond relevance retrieval. J. Am. Soc. Inf. Sci. Technol. 58(12), 1838—1850 (Oct 2007)

[7] Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2(3), 27:1—27:27 (May 2011)

[8] Liu, B.: Sentiment analysis and subjectivity. Handbook of Natural Language Processing, pp. 627—666 (2010)

[9] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1—135 (2008)

[10] Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Sentiful: A lexicon for sentiment analysis. Affective Computing, IEEE Transactions on (99), 1—1 (2011)

[11] Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of LREC. (2010)

[12] Zhao, J., Dong, L., Wu, J., Xu, K.: Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1528—1531. ACM (2012)