Magistry *et al. Lingua Sinica* (2016) 2:1
DOI 10.1186/s40655-015-0010-8

Lingua Sinica
a SpringerOpen Journal

RESEARCH Open Access

Sentiment detection in micro-blogs using unsupervised chunk extraction

Pierre Magistry, Shu-Kai Hsieh and Yu-Yun Chang*

*Correspondence:
yuyun.unita@gmail.com
Graduate Institute of Linguistics, National Taiwan University, Taipei City, Taiwan

Abstract
In this paper, we present a proposed system designed for sentiment detection for micro-blog data in Chinese. Our system surprisingly benefits from the lack of word boundary in Chinese writing system and shifts the focus directly to larger and more relevant chunks. We use an unsupervised Chinese word segmentation system and binomial test to extract specific and endogenous lexicon chunks from the training corpus. We combine the lexicon chunks with other external resources to train a maximum entropy model for document classification. With this method, we obtained an averaged F1 score of 87.2 which outperforms the state-of-the-art approach based on the released data in the second SocialNLP shared task.

Keywords: Sentiment analysis, Emotion lexicon, Unsupervised learning

1 Background

Recently, due to its great potential applications such as opinion mining and topic detection, sentiment analysis on *micro-blog* data has gained much attention than ever before. The state-of-the art approaches to sentiment analysis/detection involves attributing a *polarity* to a textual message. The polarity may accept different sets of values depending on the tasks, such as ratings and binary or ternary values (positive, negative, neutral).

The original form of this work had been prepared for the participation in the shared task at the second SocialNLP workshop. The task targets on sentiment detection in Chinese micro-blogs, which posts were extracted from Plurk online service and were mainly written in Modern Standard Chinese (MSC) with some code switching or code mixing in English, Japanese, and Taiwanese Hokkien. Messages are provided with meta-data, including timestamps, user IDs of the original posters and repliers. Besides, the posts were grouped topic-wise into 95 files by the task organizers[a].

In this task, the provided Plurk micro-blogging messages are typically short and are manually annotated with positive or negative polarities by the organizer. In addition to the provided data, external resources of other kinds are also combined for this task, which will be detailed in Section 3. Although it is worth noting that applying result comparisons on different languages or corpora are hazardous for this task, we achieved a score which resembles state-of-the-art on similar tasks in other languages.

2 Related works

Recent years have witnessed an increasing interest in the domain of micro-blogs such as *Twitter*, *Facebook*, and *Sina Weibo*. Since micro-blogs possess unique characteristics of limited length, ambiguous/error-prone, rich of acronyms, slangs, and flexible and novel usages, these would hinder the exploitation of sentiment analysis techniques (Liu 2012) in general. Additionally, different genres that appear in the social media (such as *blogs, message boards, news, and micro-blogs*) have brought different levels of difficulties for the analysts. To date, there has been little work on sentiment detection using Chinese micro-blogs. In the study by Yang (2009), F1 score of 64.8 % was reported by tracing a single Plurker's post sentiments for a few months; while 77.4–81.6 % was presented in Huang et al. (2012) by applying hybrid methods using posts from Sina Weibo.

Methodologically, sentiment detection is typically conducted using a supervised learning approach. As sentiment analysis is quite domain dependent, it is very hard to come up with a generic, one-box-fits-all approach for the task. Previous works rely on a large variety of machine learning algorithms, such as support vector machine (SVM) (Hu et al. 2007), maximum entropy (MaxEnt) (Lee and Renganathan 2011), and recurrent neural network (RNN) (Socher et al. 2013). Feature selections for these algorithms are mostly bag-of-word approaches; whereas grammar-specific rules may be involved as well to refine the features (Lee and Renganathan 2011; Thelwall et al. 2010). More recently, in addition to using lexicon-based features, some practitioners advocate applying Multi-Word Expression (MWE) recognition and/or syntactic parsing (Socher et al. 2013; Hou and Chang 2013).

2.1 Our strategy

Prior works as previously reviewed, either by exploiting various methods or resources, such as adding supplementary features, enhancing text preprocessing steps, and performing different weighting schemes onto the words, to some extent, all of these stick to the extension and improvement of bag-of-word models. Therefore, in this paper, we would take bag-of-word approaches to construct feature sets for sentiment detection.

However, some concerns exist when using bag-of-word models. In Chinese, since there are no delimiters used in word boundaries, a long controversy over wordhood issues has been raised among linguists. While applying bag-of-word models, the existence of wordhood needs to be assumed in advance, and a word segmentation system should be involved in preprocessing steps, for languages whose orthographic word boundary is not explicitly marked. Computational works in Chinese NLP-related studies used to take the task of wordhood assessment or word segmentation as a discrete (binary) decision, instead of continuous data segmentation, and have to presume a priori agreement (in whatever sense) that guides the production of segmented data, which could be further divided into training and testing data for follow-up procedures.

In our strategy, to begin with a bag-of-word approach applied to Chinese, as similar to those proposed for languages (e.g., English) written in Latin-based scripts, a Chinese Word Segmentation (CWS) procedure would be needed at the very beginning of the processing. Although this procedure is very likely to add some noisy labels to the data, but a naive tokenization on characters without any other segmentation procedure (i.e., a bag-of-character approach) is very unlikely to succeed in Chinese. Therefore, we have introduced an unsupervised CWS system into this task, regardless of the lack of training corpora specific to micro-blogs for training a CWS system. However, it is also found that segmented data resulting from CWS does not necessarily carry emotional concepts if not being placed into context. Thus, we argue that it would be more sensible to aim directly at larger chunks, which might correspond to MWEs in other languages.

Systems which do not follow the bag-of-word approach typically rely on syntactic parsing to obtain larger constituents and base their analysis on more complex structures. Unfortunately, such approaches are problematic to address Chinese micro-blog classification as they rely on current segmentation, tagging, and syntactic parsing systems which are trained on a very different genre of texts, which is typically cleaner and more formal. Given the lively and spontaneous nature of micro-blogs, it is hazardous to rely on processing tools which were trained on mostly journalistic data. On top of this, state-of-the art methods for English rely on a sentiment Treebank in which each node of the syntactic tree is annotated. Such resources are only available for English and would be time consuming to develop one for Chinese or other languages. Moreover, given how the topic can affect the polarity lexicon, the adaptability of the method to other domain is questionable. Hence, our choice is to rely on an unsupervised analysis which only needs raw data to build the features for our supervised classifier.

3 Methods

We believe that joining this task using micro-blogs would also be a very good test case for our unsupervised segmentation system (Magistry and Sagot 2012). In our segmentation system, MWE-like chunks are extracted instead of words, and the polarity of each chunk is directly learned from the training data. This method turns out to be particularly efficient as relevant keywords or key expressions for discriminating polarities within context may well be corpus specific. Therefore, in this case, the chunks with automatically learned polarities should be used in this task only but not getting collected into generic sentiment lexica lists. As data sparsity and dynamics remain a challenging issue under this approach, despite the extracted chunks, we also tried to enrich the lexica using other sources. In this section, materials and detailed procedures of the study are presented below.

3.1 Datasets for training and testing

In this task, the organizers collected 95 topics from Plurk, with 64,410 responses (containing 90 topics) for training and 6.976 responses (with 95 topics) for testing. Each messages is manually annotated with a positive or a negative value. Despite the provided Plurk training data, to address data sparsity and the OOV (out of vocabulary) problem, we rely on using various external resources. Four different sources of relevant data to help build sentiment lexica are selected. The external resources contained are Plurk emotion corpus (Chen et al. 2010), emotion words from the study on Standard Stimuli and Normative Responses of Emotions (SSNRE) in Taiwan (Cheng et al. 2012), PTT 黑特板 *heite ban* “Hate” and 黑皮板 *heipi ban* “Happy” boards, and the MOE Dictionary[b]. Figure 1 shows all of the training data that are included in this task. Samples of training data (Example 1) and testing data (Example 2) are provided below.

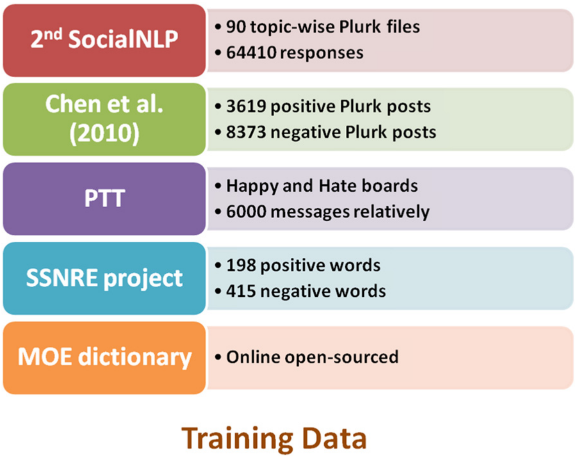

Fig. 1 The training datasets used in this task

- Training data tagged as positive

(1) 聽到時覺得很火 . . .
tingdao__shi__juede__hen__huoda
heard__SHI__feel__very__angry
When heard (the news), (I) feel very angry. . .

- Testing data:

(2) 把人民當傻瓜耍的爛政府 . . .
ba__renmin__dang__shagua__shua__de__lan__zhengfu
BA__people__DANG__fool__play__DE__bad__government
The bad government that sees people as fools. . .

In previous mood classification experiments by Chen et al. (2010), a set of Plurk corpus was collected and manually annotated with four emotion tags (positive, interrogative, negative, and unknown). From the corpus, a total of 3619 positive and 8373 negative posts are taken as gold standard data for training the system.

Additionally, we crawled some data from PTT to build another lexicon. PTT is the largest bulletin board system (BBS) in Taiwan, containing sheer numbers of boards. Among these boards, the Hate and Happy boards are filled with emotion-related expressions. The Hate board provides a platform for people to vent the hateful feelings of the unpleasant things they encounter daily; whereas, Happy board collects all the delightful things that people experience everyday. We randomly retrieved 6000 messages from the two boards relatively. Those who posted on the Hate board are considered to be negative, and those who posted on the Happy board are considered to be positive.

In the SSNRE project, emotion words could be grouped into *emotion-inducing* and *emotion-describing* words. Additionally, a total of 395 *emotion-inducing* words and 218 *emotion-describing* words have underwent three psychological experiments with a 9-point Likert scale. Based on the values evaluated via experiments, the *emotion-inducing* words (140 positive and 255 negative words) and *emotion-describing* words (58 positive and 160 negative words) could be further categorized into positive and negative word lists for training the system.

Moreover, MOE Dictionary, an online open-sourced Chinese dictionary, is also collected as training data.

Therefore, a total of five different sources of data are prepared for training the sentiment classifier using MaxEnt modeling, and the system is to be tested on the given Plurk testing data, as shown in Fig. 2.

With these training data, there are three more steps to process before feeding these data into a MaxEnt classifier.

- Unsupervised chunk extraction using unsupervised CWS system (with NVBE formulation)
- Supervised polarity identification of extracted chunks (with a binomial test)
- Supervised sentiment detection of micro-blog messages (with a Maximum Entropy model)

Details regarding the above three steps are illustrated in the following sections, and a flowchart of the procedures could be referred in Fig. 3.

3.2 Unsupervised extraction of the chunk list

To begin with, the provided Plurk training data, Chen et al. (2010) Plurk data, and PTT corpus are taken to train the unsupervised CWS system (Magistry and Sagot 2012). Within the unsupervised CWS system, Magistry and Sagot (2012) have proposed a state-of-the-art unsupervised segmentation algorithm, based on an autonomy measure computed from a normalized variation of branching entropy (NVBE). After the above unsupervised CWS system is trained, NVBE algorithm is then applied to extract a list of chunks. Here is a brief recall of NVBE formulation.

Given an n gram $x_{0..n} = x_{0..1}\, x_{1..2} \ldots x_{n-1..n}$ (where the indices are indexing the positions between every two characters) with the set of possible right contexts $\chi_{\rightarrow}$, we define its *right branching entropy* (RBE) as:

$$\begin{aligned}\mathrm{RBE}(x_{0..n}) &= H(\chi_{\rightarrow} \mid x_{0..n})\\ &= -\sum_{x\in\chi_{\rightarrow}} P(x \mid x_{0..n}) \log P(x \mid x_{0..n}).\end{aligned}$$

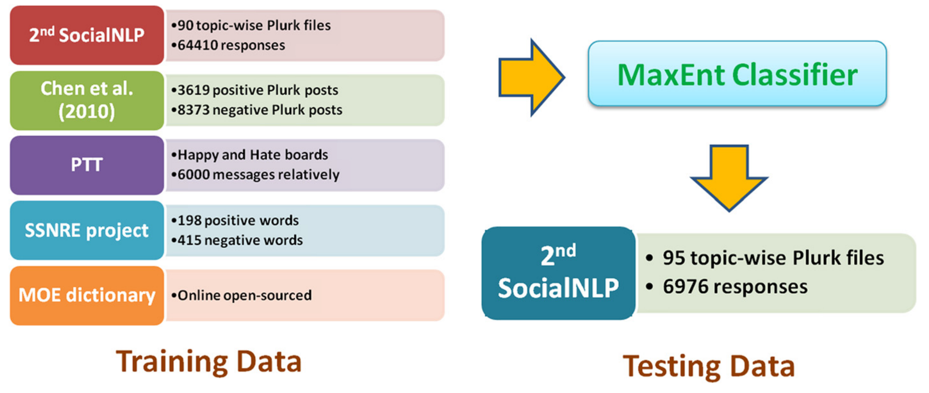

Fig. 2 The basic idea of the training and testing datasets for a MaxEnt classifier

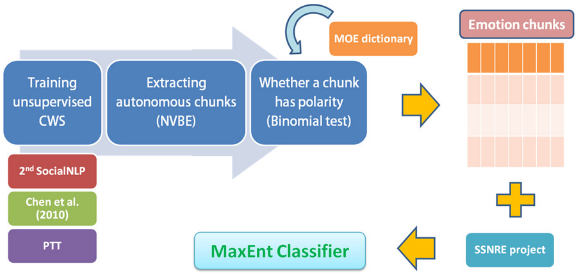
MOE dictionary
Emotion chunks
Training unsupervised CWS
Extracting autonomous chunks (NVBE)
Whether a chunk has polarity (Binomial test)
2nd SocialNLP
Chen et al. (2010)
PTT
MaxEnt Classifier
SSNRE project

Fig. 3 A detailed procedure of training a MaxEnt classifier

The *left branching entropy* (LBE) is defined in a symmetric way: if we note $\chi_{\leftarrow}$ the set of possible left contexts of $x_{0..n}$, its LBE is defined as:

$$\mathrm{LBE}(x_{0..n}) = H(\chi_{\leftarrow} \mid x_{0..n}).$$

The $\mathrm{RBE}(x_{0..n})$ can be described as $x_{0..n}$'s *branching entropy* (BE) when reading from left to right, whereas the LBE is $x_{0..n}$'s BE when reading from right to left. From $\mathrm{RBE}(x_{0..n})$ and $\mathrm{RBE}(x_{0..n-1})$ on the one hand and from $\mathrm{LBE}(x_{0..n})$ and $\mathrm{LBE}(x_{1..n})$ on the other hand, we estimate the VBE in both directions, defined as followed:

$$\begin{aligned}\mathrm{VBE}_{\rightarrow}(x_{0..n}) &= h_{\rightarrow}(x_{0..n}) - h_{\rightarrow}(x_{0..n-1})\\ \mathrm{VBE}_{\leftarrow}(x_{0..n}) &= h_{\leftarrow}(x_{0..n}) - h_{\leftarrow}(x_{1..n}).\end{aligned}$$

To allow meaningful comparison of strings with different lengths, a normalization procedure is required. Using the z-score function, NVBE is defined below.

$$\mathrm{NVBE}(x) = \frac{\mathrm{VBE}(x) - \mu_k}{\sigma_k},$$

where μ_k is the mean of all the values $\mathrm{VBE}(y)$ such as $\mathrm{len}(y) = k$ and σ_k is the standard deviation of all the values $\mathrm{VBE}(y)$ such as $\mathrm{len}(y) = k$.

In previous works, NVBE has shown to be a good empirical estimate of syntactic autonomy of a n gram. The segmentation was performed by maximizing the autonomy estimate. However, for our task, we do not need the actual segmentation. We simply extract a list of autonomous expressions (possibly larger than words) in which their polarity is tested afterwards. For our task, we do not need to actually perform the whole segmentation procedure (i.e., without outputting/using the final segmented results) but to train and use the algorithms from the proposed unsupervised CWS system, in order to get a list of chunks for the follow-up polarity testing. We select an autonomous chunk based on NVBE values, with positive values presented on the left and right branchings.

One thing to be noted is that although punctuation marks and non-Chinese characters are usually handled or removed during pre-processing steps, this kind of information remains in this task. This is done for the reason that in micro-blog messages, information such as punctuation marks and other non-Chinese characters can be used in surprising ways. For example, smileys, frequently used in micro-blogs for conveying emotions, are mostly composed of punctuation marks, symbols, and characters. Therefore, with this characteristic of micro-blogs, we expect to recognize smileys in an unsupervised fashion.

3.3 Constructing endogenous polarity chunks

Once a list of autonomous chunks is extracted, we need to determine whether these chunks are specific to a polarity or not. Here, the resource of MOE dictionary is also included to be identified and enlarge the chunk list. In order to do that, the number of times they appear in the training data containing positive and negative messages are counted, and then are compared with the overall proportion of positive over negative messages using a binomial test.

Here, the obtained p value from a binomial test is not identified in a traditional way which a significance level at 0.05 or 0.01 are usually defined. As we may want to favor large coverage of chunks over statistical significance, the scope is broader. However, to keep some information about the level of significance or certainty regarding our decisions, we use binning to define 5 classes of positive and 5 classes of negative items. We call these 5 classes of *confidence levels* (in our experiment, $0.3 > A > 0.2 > B > 0.1 > C > 0.05 > D > 0.01 > E$) and use them to further construct a list of autonomous chunks with endogenous polarities (in short, named as endogenous polarity chunks).

3.4 Maximum entropy classification

As the approach proposed in this paper is based on lexicalized features, we combine lexica from multiple sources and expect the features to be interdependent with a large amount. Additionally, it is hoped that once trained, the model can provide insights of the lexica and the task. For these reasons, we prefer to choose MaxEnt modeling as our machine learning algorithm for sentiment detection over other models, which are biased by having overlapping features (e.g., Naive Bayes models) or are used as uninterpretable black boxes (e.g., SVM or RNN).

Since the lexica from SSNRE project are already tagged with polarities through psychological experiments, this resource does not need to be involved in the above two steps (calculated by NVBE algorithms and binomial test) but to be taken directly as features for training a MaxEnt classifier. So far, the resulting lexica and chunks to be used in the classifier could be summarized in Table 1.

Despite this resource, in order to build a more complete feature set from each message, the following procedures are considered for the generated endogenous polarity chunks:

1. Use the endogenous polarity chunks to perform a maximum forward matching segmentation on the given Plurk training dataset, looking for chunks with polarities of higher confidence levels and of the longest length first.
2. Determine and annotate whether the polarity of a chunk is mostly positive or negative.

Table 1 Summary of the final lexica and chunks prepared for training a MaxEnt classifier

Source	Wordlist extraction	Polarity classification	#Positive	#Negative
SSNRE	Manual	Manual (exogenous)	198	415
MOE dictionary	Manual	Automatic (endogenous)	1370	9297
Provided dataset	Unsupervised	Automatic (endogenous)	11,692	17,950
Chen et al. (2010)	Unsupervised	Automatic (exogenous)	4867	7367
PTT Hate and Happy	Unsupervised	Automatic (exogenous)	5709	21,020

3. Extract positive and negative chunk lists from the segmented data to form a feature set prepared for sentiment classification using maximum entropy modeling. Examples are as below:
    - Positive chunk: 掌聲 鼓勵 鼓勵 *zhangsheng guli guli* “applause”
    - Negative chunk: 哇哩勒 *walile* “what the hell”

The extracted features could then be used to tell whether a larger portion of the message is covered by positive or negative chunks. Example 3 presents an example of a post with its polarity tagging and extracted features. Despite the fact that this post contains mostly negative words based on the binomial tests, the features based on the positive words are given more weights and the message is finally classified as positive[c].

(3) ㄎㄅ這句笑到我淚都流出來了你們是多喜歡香蕉阿
kaobei__zheju__xiaodaowo__lei__dou__liu__chulai__le__nimen__shiduo __xihuan__xiangjiao__a
damn__this-sentence__make-me-laugh__tear__all__flow__come-out__LE__you__how-much__like__banana__AH
-(D)__+(C)__+(B)__X__-(B)__-(E)__-(E)__-(B)__-(B)__-(E)__+(D)__+(E)__X
Damn. This sentence is so funny that I am laughing my ass off. How could you guys be so into bananas?

4 Results and discussion

The testing results from the organizer of shared task, we achieved an averaged F1 score of 87.2 %, where random guess is 54.5 % and all positive baselines are 78.19 %. Considering the possible improvements which will be listed in the next section, we would argue that our system already provides a fairly good solution to the task and can serve as a good basis to be improved on in order to reach the inter-annotator agreement rated with a little effort. Specifically, via per-topic evaluation, the results show that our systems performed unequally on various topics. Though the mean value of F1 score is 87.2 %, the standard deviation value is 12.3 %. Moreover, through the testing data, we observed that the F1 score ranges from 100 to 41.3 among the 95 topics, see Fig. 4. Within the topics where our system performed badly, we found that this is due to the inconsistencies of polarity annotation on the same or very similar messages. For example, the message 早安 *zaoan*

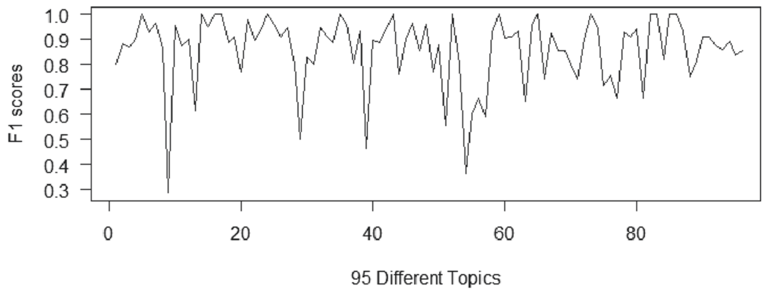

Fig. 4 The F1 scores of 95 different topics

"good morning" and 有趣 中肯 温馨 *youqu zhongken wenxin* "interesting, right to the point, warm" in some topics are mostly tagged as negative; while in others, they are annotated as positive. It is still unclear whether this is due to annotation inconsistencies between annotators or the messages imply ironic information that should have been detected with the use of meta-data.

Due to time constraints of the shared task, many options in the development of our system were left uncovered. These include:

- The use of the auxiliary data as training for the MaxEnt model
- One more step to rebuild the unsupervised lexica/chunks with insights from the model's feature weights
- Taking advantage of the division into topics and of the provided meta-data

Future refinement of the system could also include rules for feature modification as has been done in Lee and Renganathan (2011). Since some topics only contain a small amount of messages, we decided to leave the topic division aside in our first experiment. This study shows that we have not reached the plateau of the learning curve yet. However, the final evaluation results present that our model can still benefit from a larger training data. In addition, the ablation test showing how much lexical resources contribute to the score will be conducted as well.

5 Conclusions

So far, most of the works on Chinese sentiment analysis have been heavily relied on *bag-of-word*-alike paradigm which requires identifying *wordhood* as the pre-processing task. On the contrary, we propose an unsupervised CWS model into sentiment analysis. We believe that the model proposed in this paper is very promising not only due to its treatment of necessary indeterminacy of wordhood issue with agnostic methodology but also its great achievement of performance over others in sentiment-related processing tasks.

Endnotes

[a] Unfortunately, participants in this shared task were too few to provide a comprehensive comparison of the results.

[b] https://www.moedict.tw/about.html.

[c] It is found that there was a popular and funny banana dance video widely spread among the Plurkers at that time, which explains why banana is tagged as positive with the highest confidence level in this message.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors were involved in the works of data processing, system construction and manuscript preparation. PM worked on building the system architecture; YC collected the data and analyzed the results; and SH prepared the scope of the manuscript in cooperation with all authors. The final manuscript was read and approved by all authors.

Acknowledgements

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2010-5094-7. Additionally, we thank the anonymous reviewers' comments on this study.

Received: 30 July 2015 Accepted: 26 December 2015
Published online: 26 January 2016

References

Chen, Mei-Yu, Hsin-Ni Lin, Chang-An Shih, Yen-Ching Hsu, Pei-Yu Hsu, and Shu-Kai Hsieh. 2010. Classifying mood in plurks. In Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), 172–183. Puli, Nantou, Taiwan: The Association for Computational Linguistics and Chinese Language Processing.

Cheng, Chao-Ming, Hsueh-Chih Chen, and Shu-Ling Cho. 2012. Affective words. In A study on standard stimuli and normative responses of emotion in Taiwan. Taipei, Taiwan: National Taiwan University.

Hou, Wen-Juan, and Chuang-Ping Chang. 2013. Sentiment classification for movie reviews in Chinese using parsing-based methods. In International Joint Conference on Natural Language Processing, 561–569. Nagoya, Japan: Asian Federation of Natural Language Processing.

Hu, Yi, Ruzhan Lu, Yuquan Chen, and Jianyong Duan. 2007. Using a generative model for sentiment analysis. Computational Linguistics and Chinese Language Processing 12: 107–126.

Huang, Sui, Jianping You, Hongxian Zhang, and Wei Zhou. 2012. Sentiment analysis of Chinese micro-blog using semantic sentiment space model. In 2012 2nd International Conference on Computer Science and Network Technology (ICCSNT) 1443–1447. Changchun, China: IEEE.

Lee, Huey Yee, and Hemnaath Renganathan. 2011. Chinese sentiment analysis using maximum entropy. In Proceedings of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP), IJCNLP 2011, 89–93. Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

Liu, Bing. 2012. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies 5: 1–167.

Magistry, Pierre, and Benoît Sagot. 2012. Unsupervized word segmentation: The case for Mandarin Chinese. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, 383–387. Jeju, Republic of Korea: Association for Computational Linguistics.

Socher, Richard, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Conference on Empirical Methods in Natural Language Processing (EMNLP) 1631–1642. Seattle, USA: Association for Computational Linguistics.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Cai Di, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61: 2544–2558.

Yang, Ting-Hao.2009. Mining blogger's glossary impressions from a micro-blog corpus. Master thesis. Hsin-Chu Taiwan: National Tsing-Hua University.