Sentiment Analysis Based on User Tags for Traditional Chinese Medicine in Weibo

Junhui Shen[1], Peiyan Zhu[2], Rui Fan[2], Wei Tan[3], and Xueyan Zhan[4(✉)]

[1] Information Center, Beijing University of Chinese Medicine, Beijing 100029, China
[2] State Key Lab of Software Development Environment, Beihang University, Beijing 100191, China
[3] School of Management, Beijing University of Chinese Medicine, Beijing 100029, China
[4] School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing 100029, China
zhan_xueyan@163.com

Abstract. With Western culture and science been widely accepted in China, Traditional Chinese Medicine (TCM) has become a controversial issue. So, it is important to study the public's sentiment and opinions on TCM. The rapid development of online social network, such as twitter, make it convenient and efficient to sample hundreds of millions of people for the aforementioned sentiment study. To the best of our knowledge, the present work is the first attempt that applies sentiment analysis to the fields of TCM on Sina Weibo (a twitter-like microblogging service in China). In our work, firstly, we collected tweets topics about TCM from Sina Weibo, and labelled the tweets as supporting TCM or opposing TCM automatically based on user tags. Then, a Support Vector Machine classifier was built to predict the sentiment of TCM tweets without tags. Finally, we presented a method to adjust the classifier results. The performance of F-measure attained by our method is 97 %.

Keywords: Sentiment analysis · Machine learning · Support Vector Machine · Traditional Chinese Medicine (TCM) · Weibo · User tag

1 Introduction

At the beginning of the 20th century, western culture and science was introduced in China and soon was accepted. Traditional Chinese Medicine (TCM) was seriously thrown into doubts in terms of its scientific foundation. When such kind of debates are reviewed in respective of debaters' sentiment towards TCM, two types of sentiment are dominating: one school thinks that TCM cannot be proved by scientific experiments, so it is pseudo-science and should be abolished, while the other school believes TCM is effective in treating many diseases, therefore TCM is essentially a kind of science.

Microblogging today has become a very popular communication tool among internet users. In China, Sina Weibo (http://www.weibo.com), a Twitter-like

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 134–145, 2015.
DOI: 10.1007/978-3-319-25207-0_12

microblogging service launched in 2009, has accumulated more than 500 million users in less than four years, leading to its most important role in the social media marketing platform. Every second, approximately more than 1000 Chinese tweets are posted in Weibo. It is imaginable that the debates about TCM spread into cyber-space in an unbelievable speed.

So far, although many researches have been conducted on sentiment classification [1], there is little such work about Traditional Chinese Medicine. To the best of our knowledge, the present work is the first attempt that applies sentiment analysis to the fields of TCM on Sina Weibo. In our work, main contents are: collecting corpus and dictionary resources, labelling data automatically based on user tags, building a Support Vector Machine (SVM) classifier to predict the sentiment of TCM tweets and presenting a method to adjust the classifier results. The performance of F-measure attained by our method is 97%.

2 Data Collecting and Labelling

In this section, we discuss the collection and pre-processing of tweets which topics are about TCM. For each tweet in our corpus, we converted it into a sequence of words.

2.1 Corpus Collection Based on User Tags

In China, Sina Weibo is one of the most important social networking channels, and is the Chinese counterpart to Twitter. As with Twitter, Weibo users are allowed to post real-time messages, called tweets. Tweets are short messages, restricted to 140 characters in length.

There are some prominent differences between Twitter and Weibo. For example, users can freely tag himself/herself to indicate his/her interests and characteristics in Weibo. Of course, tagging is not mandatory in Weibo where users can tag up to 10 keywords.

In January 2014, we searched Weibo users interested in TCM by user tags. If someone has more than one user tags included in our search keywords list, he/she would be duplicated in our dataset. After filtering the duplicated users, we constructed a dataset including 48861 Weibo users, denoted as C. The user tags and the corresponding number of Weibo users are listed in Table 1. Among all tags, “Traditional Chinese Medicine” is used by 42608 users and occupies the dominating share of 87%, “Medicine Material”, “Acupuncture and Moxibustion” and “Massage” follow but none of them takes the share of more than 8%. It is not surprised to find that “Traditional Chinese Medicine” is the main tag used because it is a wide concept, which often refers to not only TCM therapy but also includes “Medicine Material”, “Acupuncture and Moxibustion” and “Massage”. Using the Application Programming Interfaces (API) provided by Weibo, we collected the tweets which were posted by the users in C. Due to the limit of API, only the most recent 2000 tweets each user posted can be obtained, we gathered 21,242,370 tweets totally.

Table 1. The user tags and the corresponding number of Weibo users.

User Tag(Original Text)	User Tag(Translation)	the Counts of Weibo Users
中医	Traditional Chinese Medicine	42608
中药	Medicine Material	3827
针灸	Acupuncture and Moxibustion	3236
推拿	Massage	2198
艾灸	Moxa-moxibustion	763
中草药	Chinese Herb Medicine	417
针刺	Acupuncture	73
针推	Acupuncture and Massage	67
中成药	Chinese Patent Drug	50

The sentiment of a retweet is not always consistent with the tweet, especially when debating. For this reason, we split each tweet which followed with retweet and inserted each retweet into our corpus. Sometimes, one post had more than one re-posting, so we had much more tweets after splitting. Totally, we collected 43,012,068 tweets in our corpus, more than twice of original tweets amount.

2.2 Two Dictionary Resources

In this paper, we introduced two new resources for the pre-processing of Weibo data topics on TCM: custom dictionary and TCM terminology dictionary. We collected western medicine terminology, TCM terminology and popular vocabulary on the internet, totally 5307 words in the custom dictionary. It can be used as a helpful complement of built-in dictionary of general tool for Chinese Word Segmentation. The TCM terminology dictionary contained 2715 TCM terminology words including Traditional Chinese Medicine, Chinese Patent Medicine, Chinese Herb Medicine and acupuncture point etc. It can be used to filter the Weibo which topic is about TCM.

2.3 Pre-processing of Data

We pre-processed all the tweets as follows:

1) Translating the tweet to Chinese Simplified if it is written by Chinese Traditional;

2) Filtering URL links (e.g. http:// example.com), Weibo user names (e.g. @shenjh-with symbol @ indicating a user name), Weibo special words (e.g. reply), and emoticons from tweets;

3) Segmenting Chinese Word (with the ICTCLAS tool [2] and the custom dictionary as introduced in section 2.2) to generate a sequence of words;

4) Removing stop words (such as “oh”) from the bag of words;

5) Filtering advertisements by key words (such as “sale”).

2.4 Filtering Chinese Medicine Tweets

However, the topics of tweets posted by the users interested in TCM were diverse and not only about TCM. Therefore, in our study, we should screen out the tweets in which the real topic was not about TCM.

During our process, we filtered the tweets topics on TCM with the TCM terminology dictionary (introduced in section 2.2). Usually, a tweet on TCM contains a few key words which are about TCM, so we filtered the tweets including at least two different key words of TCM strictly. After filtering, 1,650,497 tweets remained in our corpus in which the real topic is about TCM.

2.5 Labelling the Data

When we were labelling the sentiment of tweets, our approach based on the basic principle: the user prone to have consistent opinions for a certain topic due to the principle of consistency [3]. It means that if the user's opinion is for TCM, the sentiment of all the tweets he/she posted is for TCM. In contrast, if the user's opinion is against TCM, the sentiment of all the tweets he/she posted is against TCM.

In our analysis, we acquired user's opinions about TCM by the user tags. The key words used as user tags are defined by the user. Consequently, the user tags could be different even if the sentiment to TCM is same. The user tags used to label the sentiment are listed in Table 2. Only the user tags which had been quoted by more than 10 users are included in the table. As a result, 1866 Weibo users were labelled as supporting TCM, while 290 Weibo users were labelled as opposing TCM. The rest were not labelled because we couldn't obtain obvious sentiment orientation from his/her user tags. Based on our basic principle, we labelled the sentiment of tweets according to the user's opinions on TCM. Finally, 40888 tweets were labelled as supporting TCM, and 6975 tweets were labelled as opposing TCM. Obviously, there was an imbalance but it is consistent with the reality. The tweets labelled would be used as the training dataset in the next step of our research.

3 Methodology

This section presents the methodology of sentiment classification system we used. First, feature selection method was used to pick out discriminating terms for training and classification. Then we used the machine learning method to build a sentiment classifier. Finally, we adjusted the classification results based on the basic principle that a user keeps consistent opinions for a certain topic.

3.1 Feature Selection

A number of feature selection metrics had been explored in text categorization, i.e. chi-square (CHI), information gain (IG), correlation coefficient (CC) and

Table 2. The user tags and corresponding Weibo user counts.

Sentiment	User Tag(Original Text)	User Tag(Translation)	User Counts
Supporting TCM	中医爱好	Love TCM	972
	爱中医	Love TCM	239
	中医师	Doctor of TCM	230
	喜欢中医	Love TCM	85
	中医粉	TCM Follower	55
	中医控	TCM Follower	52
	中药师	Pharmacist of TCM	51
	针灸师	Acupuncturist	42
	中医养生爱好	Regimen of TCM	29
	推拿师	Masseur	28
	中医达人	TCM Master	12
Opposing TCM	反中医	Oppose TCM	191
	中医黑	Abominate TCM	55
	反对中医	Oppose TCM	28

odds ratios (OR). All these methods computed a score for each individual feature and then picked out a predefined size of a feature set. In our approach, we used chi-square feature selection method, one of the most effective methods in text categorization [4]. Chi-square measures the lack of independence between a term t and a category ci and can be compared to the chi-square distribution with one degree of freedom to judge extremeness.

3.2 Machine Learning Method

So far, most of the researches on sentiment classification focused on training machine learning algorithms to classify reviews [5],[6]. Support Vector Machine has been shown to be highly effective for traditional text categorization [7]. Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set.

Here we limit our discussion to linear SVM due to its popularity and high performance in text categorization [8].

3.3 Adjusting Sentiment Classification Results

Based on the basic principle that the same user should have consistent opinions for a certain topic [3], we adjusted the sentiment classification results: assigned majority sentiment label to all the tweets the same user posted.

Based on the sentiment classification result, the number of tweets which are judged as supporting TCM posted by one user can be obtained as C_s, and the number of tweets posted by the same user which are judged as opposing TCM can be obtained as C_o. Then we define γ as

$$\gamma = \frac{max\{C_s, C_o\}}{C_s + C_o} \tag{1}$$

where $0.5 \leq \gamma \leq 1$. If $\gamma = 1$, it means the sentiment of the user is consistent absolutely. If $\gamma = 0.5$, it means Co is equal to Cs, then we don't need to adjust the sentiment classification result. When $0.5 < \gamma < 1$, we can adjust the classification result.

4 Experiments and Results

In our dataset, there were 1,650,497 tweets in which the topic focused on TCM, including 40,888 tweets labelled as supporting TCM, and 6,975 tweets labelled as opposing TCM (introduced in Section 2.5). Since it was imbalanced, we focused on not only the global performance, but also the performance of each class. Therefore, we chose F1 to evaluate the classification system.

After applying CHI feature selection to tweets, for all our experiments, we used Support Vector Machine and reported 5-fold cross-validation test results.

Pang et al. [7] argued that feature presence binary value is more useful than feature frequency for the SVM classier. Therefore, we used binary value for each feature instead of feature frequency.

4.1 The Performance Measure

To evaluate the imbalanced classification system, we used the F1 measure. This measure combines recall and precision in the following way:

$$Precision = \frac{number\ of\ correct\ positive\ cpredictions}{number\ of\ positivec\ predictions} \tag{2}$$

$$Recall = \frac{number\ of\ correct\ positive\ predictions}{number\ of\ positive\ examples} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Recall + Precision} \tag{4}$$

4.2 Feature Selection Results

The top 10 key words of each class selected by the CHI method were listed in Table 3. Among the proponents of TCM, it is not surprising that "Medicine Material", "Health Preservation", "Traditional Chinese Medicine" and "Body" were often used. The frequency of "State" and "China" could be due to that

Chinese government employed clear policy to support TCM. Among the opponents of TCM, "Aristolochia acid", "Cinnabar", "Longdan Xiegan Wan" and "Injection" were popular words. This could be attributed to that all these terms are related to untoward effects so the opponents wanted to shake the scientific foundation of TCM.

Table 3. The top 10 key words of each class selected by the CHI method

Supporting TCM (Original Text)	**Supporting TCM** (Translation)	**Opposing TCM** (Original Text)	**Opposing TCM** (Translation)
中药	Medicine Material	中成药	Chinese Patent Medicine
养生	Health Preservation	马兜铃酸	Aristolochic acid
国家	State	注射	injection
科学	Science	注射液	injection
中医药	TCM	方舟子	Zhouzi Fang
中国	China	朱砂	Cinnabar
身体	Body	事件	Events
医生	Doctor	反对	Oppose
健康	Health	马兜铃	Aristolochic
治疗	Cure	龙胆泻肝丸	Longdan Xiegan Wan

Figure 1 shows the classification performance curves using the CHI feature selection method vs. feature number. The performance of classifier is above 90% and the performance increases as the number of features increases. It is found that the performance of TCM proponent classifier is slightly better than the performance of the total classifier. It is notable that the performance of TCM opponent classifier increases significantly when the number of features increases. The performance of each class is relatively stable when the number of features exceed 3000. So, we fixed the number of features at 3000 in the following experiments.

4.3 Classification Results

Because the dataset was imbalanced, we tuned the wi parameter for SVM, where 0=wi=1. Figure 2 shows the performance of F1, precision and recall by varying the parameter wi from 0.1 to 1.0. When wi increases from 0 to 1, precision, recall and F1 all increase significantly and reach plateau. Figure 3 shows precision and recall separately with each class by varying the parameter wi. It is interesting that precision shows a reverse trend of that of recall. When wi increases from 0 to 1, the precision of supporting TCM gradually decreases while the precision of

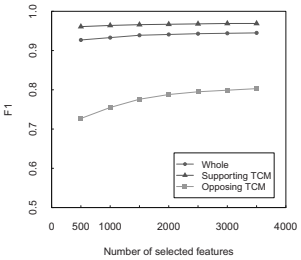

Fig. 1. The classification performance curves using the CHI feature selection method vs. feature number.

opposing TCM rapidly increases. During the same process, recall of supporting TCM increases while recall of opposing TCM significantly decreases. Figure 4 shows the performance of each class separately.

From these figures we can see that it is better to set wi to 0.9. It summarizes the performance of the classifier of supporting TCM and the classifier of opposing TCM. When wi gradually increases, for TCM proponents, Precision decreases from 98% to 96%, Recall increases from 91% to 98%, and F1 increases gradually to a plateau phase. For TCM opponents, Precision increases 62% to 86%, Recall decreases from 89% to 75%, and F1 increases gradually to a plateau phase.

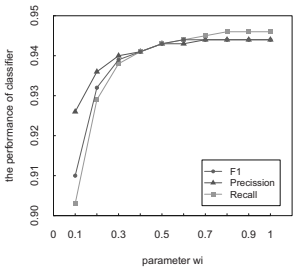

Fig. 2. The performance curves by varying the parameter wi

We viewed both the whole and the individual class: when wi increases from 0.1 to 1, F1 value increases gradually to a plateau phase. F1 value reaches the optimal when wi equals to 0.9.

4.4 Adjusted Classification Results

As introduced in Section 4.3, we can adjust the classification results based on the principle that the same user should have consistent opinions for a certain topic. Figure 5 shows the performance by varying the parameter γ from 0.5 to 1(and fixing wi=0.9). There is a notable decline of F1. When γ is set to 0.5, our model achieves the best performance of F1, which is 97%.

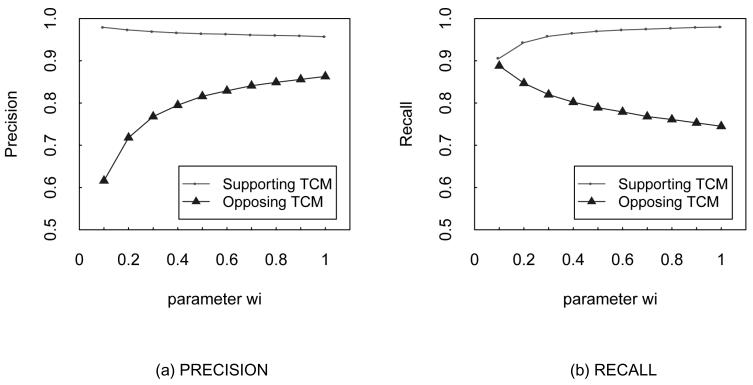

Fig. 3. The precision and recall separately with each class by varying the parameter wi.

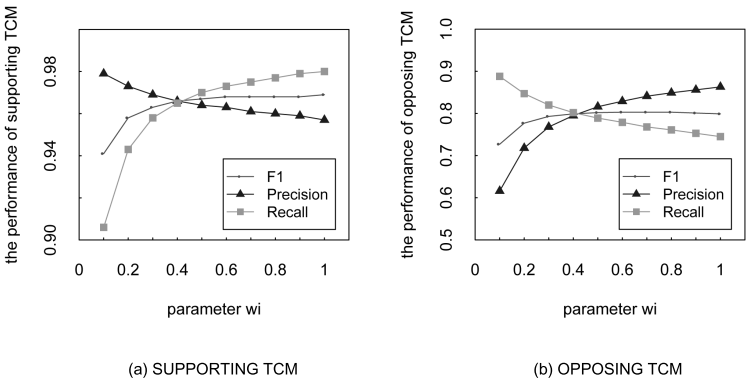

Fig. 4. The Performance of Supporting TCM and Opposing TCM Separately.

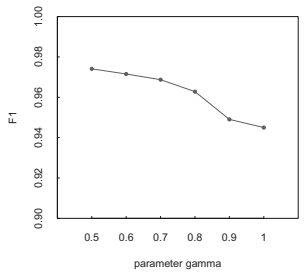

Fig. 5. The performance curve of sentiment classification by varying the parameter γ from 0.5 to 1.

4.5 Prediction

Besides the labelled tweets, there are 1,602,634 unlabelled tweets which the topic is about TCM. We can predict their sentiment with our trained classifier. Figure 6 shows the curves for the amount of tweets which, respectively, support TCM (a) and oppose TCM (b). The amount of tweets supporting TCM far exceeds the number of tweets opposing TCM. For the simple comparison, the tweets amount of both opposing and supporting TCM are converted to their log forms, as shown in (c). This result coincides with the real world. In china, most people support TCM, especially the regimen of TCM. There are only a small number of people opposing TCM. In addition, the tweets amount before 2010 was very small, due to the limit of Weibo where only the most recent 2000 tweets of each user can be obtained.

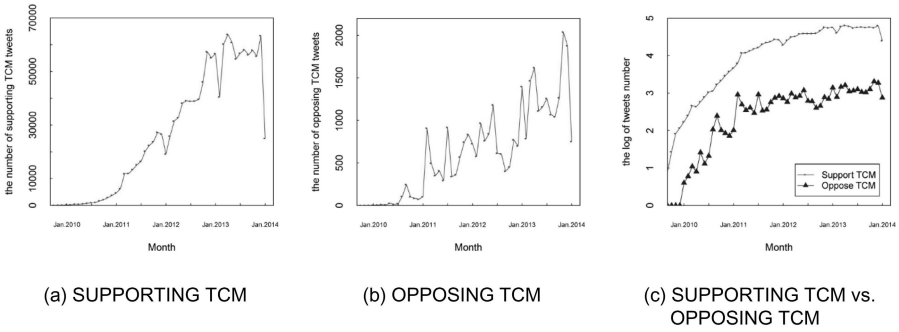

Fig. 6. The curves for the amount of tweets which, respectively, support TCM (a) and oppose TCM(b).

After the sentiment classification of tweets concerning TCM, we can monitor the sentiment fluctuation of TCM in Weibo. As shown in Figure 6, the number of tweets supporting TCM decreases significantly during January of 2012, 2013 and 2014. Because the three periods conflicted with Chinese New Year. The decrease could be due to that people did not log on Weibo during these holiday seasons. On the contrary, the number of tweets opposing TCM shows no clear trend. The erupt of the tweets opposing TCM could be caused by incidents related to TCM, which could be an interesting research topic in the future.

It is also found from Figure 6 that the tweet counts of both class reached the peak in Nov. 2013. We show the details of the curve in that month in Figure 7. In November 2013, the number of tweets supporting TCM is relatively stable while the number of tweets opposing TCM fluctuates drastically. This is in line with the overall trend of the number of each class.

Moreover, the top 50 key words of each class in Nov. 2013 is shown separately in Figure 8. “Traditional Chinese Medicine”, “Health Preservation”, “Food” etc.

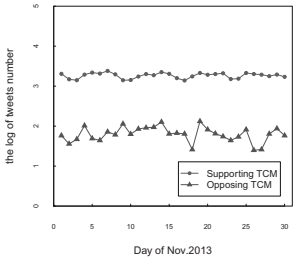

Fig. 7. The curves for the amount of tweets which, respectively, support TCM (a) and oppose TCM in Nov.2013.

often appears in tweets supporting TCM, while "Chinese Patent Medicine", "injection", "toxicity" etc. frequently appears in tweets opposing TCM. This is conformed with the Table 3. It is worth mentioning that words such as "toxic" or "harmful" appears in tweets supporting TCM too. This is not unexpected because TCM theory admits that a few TCM medicine is toxic so the dosage of these toxic TCM medicines should be controlled and be paid special attention.

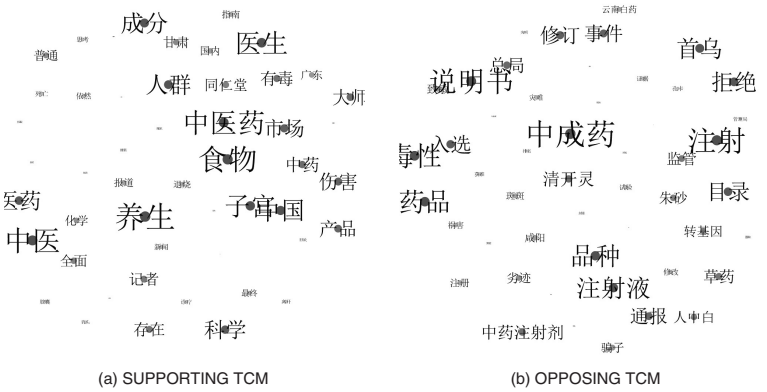

Fig. 8. The top 50 key words of each class(supporting TCM and opposing TCM) in Nov.2013.

5 Conclusion and Future Work

Traditional Chinese Medicine is an ancient but thriving and somewhat controversial discipline. Meanwhile, it is important to study the public's sentiment and opinions on TCM. To the best of our knowledge, the present work is the first attempt to study sentiment analysis for TCM based on user tags in Weibo. We

classify the opinions on TCM into two categories: supporting TCM and opposing TCM. The F1 measure value of our method is 97%.

Moreover, we collect 48861 Weibo users who are interested in TCM and 1,650,497 tweets concerning about TCM. And we construct two dictionary resources for processing Chinese tweets topic on TCM. Based on the aforementioned corpora and resources, we build an effective classifier with SVM to analyse the sentiment opinions on TCM using Weibo tweets automatically.

In future work, we will explore more linguistic techniques to study sentiment analysis for TCM, such as parsing, semantic analysis and topic modelling.

Acknowledgments. This work is supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project Name:Public opinion analysis of Traditional Chinese Medicine based on sentiment analysis, Project No: 15YJCZH137).

References

1. 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 Press, Jeju island (2012)
2. Zhang, H., Yu, H., Xiong, D. Liu, Q.: HHMM-based Chinese lexical analyzer ICTCLAS. In: Proc of SIGHAN Workshop on Chinese Language Processing, pp. 758–759. ACM Press, Sapporo (2003)
3. Deng, H., Han, J., Li, H., Ji, H., Wang, H., Lu, Y.: Exploring and inferring user - user pseudo-friendship for sentiment analysis with heterogeneous networks. Statistical Analysis and Data Mining **7**, 308–321 (2014)
4. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: 14th Int'l Conf. Machine Learning, pp. 412–420. ACM Press, Nashville (1997)
5. Liu, B.: Sentiment Analysis and Opinion Mining. International Journal **5**, 1–167 (2012)
6. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data.In: Proceedings of the Workshop on Languages in Social Media, pp. 620–622. ACL Press, Portland (2011)
7. Pang, B., Lee, L. Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of Emnlp, pp. 79–86. ACL Press, Stroudsburg (2002)
8. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: a library for large linear classification. Journal of Machine Learning Research **9**, 1871–1874 (2008)