Dividing for Combination: Semi-supervised Sentiment Classification for Microblogs

Xie Songxian and Wang Ting

Department of Computer Science and technology, School of Computer,
National University of Defence Technology, Changsha, China
{xsongx,tingwang}@nudt.edu.cn
http://www.nudt.edu.cn

Abstract. Sentiment analysis is the computational study of how opinions(positive or negative) are expressed in language. It has been an important research area of Natural Language Processing in recent years. There are many challenges for sentiment analysis especially of the social media language. This paper focuses on context dependence, which has been the most challenging problem, and proposes a novel framework to solve the problem. By dividing the feature space into two parts, a general classifier on general part of features is trained on off-the-shelf labelled idiom resources, and a context classifier on context-dependent part of features is trained on random tweets retrieved from microblogs in distant supervision manner. Then a semi-supervised framework is developed to combine the general classifier and context classifier into a bootstrapping classifier. Experiments results show that the proposed framework is effective and achieves encouraging accuracy performance.

Keywords: sentiment analysis; idioms; general classifier; context classifier;microblog

1 Introduction

Sentiment analysis is the computational study of how opinions, attitudes, emotions, and perspectives are expressed in language. Sentiment analysis provide tools and techniques for extracting subjective information from large datasets and summarizing it for successive application such as Business Intelligence, Public Opinion Analysis, and Election Prediction, etc[1]. Sentiment classification, which deals with determining sentiment orientation of target text, is the classification form of sentiment analysis[2]. Although it can be viewed as a special text categorization problem, in fact sentiment classification is a more challenging task than text classification, because sentiment expressions critically depend on domains and context[3].

With the emergence and prosperity of social networks, the amount of user-generated content (UGC) has risen exponentially over the last decade, and such content is now always at our fingertips. Mircoblog is one of the most important social networks which has absorbed millions of users to update their status

(conventionally called tweets in microblog) day and night. Beyond merely displaying news and reports, the mircoblog itself is also a large platform where different opinions are presented and exchanged. Users comment, discuss, compliment, argue and complain over topics they are interested in, sharing their own feelings publicly. It has been well recognized that UGC of microblog with rich sentiment information can trigger more attention, feedback or participation, and sentiment analysis researchers have begun to pay more and more attention to microblogs content [4]. The distillation of subjective knowledge from such abundant language resources becomes an important part of applications in fields such as commerce, tourism, politics and health. But as a freely-publishing platform, users may use or create new words of abbreviations and acronyms that seldom appear in conventional text documents, for example, words like "cooooooool", OMG","-(", are intuitive and popular in microblog. Though they may provide convenience for on-line communications of users, it is difficult for computer to accurately identify the semantic meanings of these words. To make situation worse, new words may arise and old words may change their meaning continuously in tweets. The noisy quantity, informal nature and explosive vocabulary of tweets make sentiment classification of tweets a very difficult and challenging task, because the sentiment in tweets often depends on such particular expressions as emoticons, repeated letters and exclamation, etc., which get their semantic meaning only in the context of microblogs. Context dependence has been the main challenge sentiment classification of tweets must face. In this work, we have focused on the context-dependent problem of sentiment classification on microblogs.

To solve the problem, we propose a novel semi-supervised framework based on our two rational hypotheses. Firstly the problem is formulated as feature vector space model of text classification. By dividing the feature space into two parts, a general classifier on general part of features is trained on off-the-shelf labelled idiom resources, and a context classifier on context-dependent part of features is trained on tweets randomly retrieved from microblogs in distant supervision manner. Then a semi-supervised framework is developed to combine the general classifier and context classifier into a bootstrapping classifier. The experiments carried out on a Chinese microblog platform show that the framework is effective and achieves encouraging accuracy performance.

The rest of the paper is organized as follows: Related works are discussed in Section 2. Our two hypotheses are put forward in Section 3, and our framework of semi-supervised sentiment classification is described in Section 4. The results and discussions of the experiment are presented in Section 5. Finally we conclude about our work in Section 6.

2 Related Works

The task of sentiment analysis has been a popular research area for years. Previous efforts mainly focuses on reviews and news comments. Generally, in terms of methodology, rule-based approaches and machine-learning based approaches are

two major popular methods, and the machine-learning based approaches usually act as an upper bound for other methods to compare with, in that they have higher performance because of the strong generalization ability of classifiers[1, 5].

Recently, sentiment analysis researchers have begun to pay more and more attention to the massive user-generated content of social networks such as Twitter[1]. Many studies showed that the unique characteristics of Twitter can be incorporated into sentiment analysis techniques. Barbosa and Feng [6] first investigated to use a two-stage Support Vector Machine (SVM) classifier for tweets sentiment classification which proved to be robust regarding biased and noisy data. Hu et al. [7] interpreted emotional signals available in social media data for unsupervised sentiment analysis by providing a unified way to model two main categories of emotional signals: emotion indication and emotion correlation. Jiang et al. [3] focused on target-dependent Twitter sentiment classification, they proposed to improve target-dependent Twitter sentiment classification by taking both target-dependent features and related tweets into consideration. Wang et al. [8] focused their study on hashtag-level sentiment classification, they proposed a novel graph model and further improved the model using an enhanced boosting classification setting. Amir Asiaee T et al. [9] presented a cascaded classifier framework for per-tweet sentiment analysis by extracting tweets about a desired target subject, separating tweets with sentiment, and setting apart positive from negative tweets. Hu et al. [10] extracted sentiment relations between tweets based on social theories, and proposed a novel sociological approach to utilize sentiment relations between messages to facilitate sentiment classification and effectively handle noisy Twitter data. Motivated by sociological theories arguing that humans tend to have consistently biased opinions, Guerra et al. [11] addressed challenges of topic-based real-time sentiment analysis by proposing a novel transfer learning approach with a suitable source task of opinion holder bias prediction. Thelwall et al. [12, 13] designed SentiStrength, an algorithm for extracting sentiment strength from informal English text. The algorithm built on human-evaluated dictionaries for words connotated with positive or negative sentiments and exploited the grammar and spelling styles in typical microblogs. Zhao et al. [14] presented a hierarchical generative model, called user-sentiment topic model (USTM) used in social network analysis to find influential users on topic level with sentiment information.

All works above have tried to adapted their methods to the microblogs by making use of the network and language characteristics, no matter what approaches they have taken. However, in this paper, we solve the context-dependent problem from a novel perspective by separating context-independent part of features from context-dependent part of features of sentiment classification.

[1] `twitter.com`

3 Problem Formulation and Solution

3.1 Problem Formulation

Simply speaking, sentiment classification aims to classify text as predefined sentiment polarity classes(negative or positive). Formally, Given document corpus $D = \{d_1, \ldots, d_n\}$, and predefined sentiment classes $C = \{1, -1 \mid \text{positive} = 1, \text{negative} = -1\}$, the task of sentiment classification is to predict each d_i with a label c_i. To be along with text categorization, each document can be represented as a vector of features $x = \mathbb{R}^n$, where n is the size of a pre-specified feature volume V. For sentiment classification, the weight of each entry in the vector usually is often specified as binary, with weight equals to 1 for feature present in the vector and 0 for absent. Given a training dataset $X = \{x_1, \ldots, x_m\}$, a classifier can be build:

$$f : X \longrightarrow Y, Y = \{1, -1\} \ . \tag{1}$$

and employ it to predict label for an unseen instance x by computing $f(x)$, with each instance represent as a vector $x = (w_1, \ldots, w_v)$, in which w_i is the ith features weight.

3.2 Feature Space Division

In previous sentiment classification researches, there is an underlying hypothesis, which implies all features in the text vector represent the documents sentiment polarity equally. In fact, some features may only appear in specific context, while others appear across any context. As an example, an English tweet is listed as following:

@Kid_Cloudz: Happy birthday to Yessicaaaa! :D lovee you feggit wish you the best day everrrrr!!!!! @030268.

If simple "bag-of-words" features are considered, all words would be extracted as equal features to model the sentiment when computationally classifying sentiment polarity of this tweet. However with thorough considerations, we can find that such words as "@Kid_Cloudz, :D, lovee, everrrrr,!!!!!" are mainly used in the language context of microblogs, while "Happy, birthday, wish, best, thanks" are sentiment indicators of text almost in any language context. With such intuition, we proposed a feature space division hypothesis as:

Hypothesis 1. In the feature space of sentiment classification, features can be divided into two different parts:

- context-independent part, i.e. general features, which are indicators of sentiment independent of any language context;
- context-dependent part, i.e. context features, which indicate sentiment in specific language context.

Formally, for the text feature vector of a tweet $x = (w_1, \ldots, w_l, w_{l+1}, \ldots, w_v)$, it could be divided into two parts:

$$x = \begin{cases} x_g & : \text{general features} \\ x_c & : \text{context features} \end{cases} . \tag{2}$$

where $x_g = (w_1, \ldots, w_l)$ denotes the general part of feature space, while $x_c = (w_{l+1}, \ldots, w_v)$ denotes the context dependent part.

Based on Hypothesis 1, we studied following questions in this paper:

- How to identify each divided part of features, and what does it mean to the sentiment classification of tweets?
- How to model sentiment of tweet with the two divided feature sets in the machine learning framework?
- How to construct training dataset for sentiment classifiers based on the two parts of features separately?

4 Sentiment Classification Framework

Besides traditional expressions, some languages of microblogs are often considered "Mars Language" because of their obscurity. However users could still distinguish the sentiment polarity of the tweet even if some words seem strange to them. Intuitively, this kind of phenomenon may be explained by the general part of the tweet which is used to express the holistic sentiment polarity of the author across any context, and the polarity of general sentiment words are prone to be recognized by anyone independent of context knowledge. Comparably, for sentiment classification, we put forward that, the sentiment polarity of a document could sometimes still be recognized with only general part of feature space (x_g in Equation 2). That is to say, if general sentiment knowledge could be modeled, what sentiment polarity a tweet prefers for could be still classified based on such general sentiment models.

In fact, many researchers have tried to establish all kinds of sentimental ontology lexicons to represent general knowledge of humans sentiment, such as SentiWordNet[15] and General Inquiry[16] in English, Hownet[17] and NTUSD (Chinese Network Sentiment Dictionary) in Chinese[18], etc. However, some entries of these lexical resources have multiple senses with different sense representing different sentiment polarity, and the exact sense unavoidably depends on the context. It seems a rather difficult task to find a general resources to model the context independent knowledge. Actually such knowledge exists in many cases, in which one general word or combination of a few words could identify exact sentiment polarity independent of context, such as idioms and proverbs. In this section, we propose a novel semi-supervised framwork to make use of idiom resources to model the general part of tweet sentiment konwledge and combine the model with the context-dependent model to establish a complete sentiment classifier of tweet.

4.1 General Sentiment Classifier

There are many linguistic resources highly valuable for sentiment classification, of which idiom resources attract interests of this research. Idioms are common phenomena of many languages beside Chinese, such as castles in the air, a bed of thorns, bring down the house in English. The form of idioms is succinct and the meaning is penetrating, which is a quintessence part of the language. Generally speaking, the structure of idioms is fixed and cant be changed at will; idioms have semantic intactness, which is not generally the simple summation of the literal meaning of each component; idioms chiefly use metaphor, exaggerator and comparison in the rhetoric to express its real meaning; and most importantly, the sentimental orientation of idioms is independent and unchangeable under any context. There are many off-the-shelf lexical idiom resources in all kinds of languages with entries take the example form as:

castles in the air: a derogatory term, indicate the illusive things or impractical fanciness metaphorically.

In this example, the entry is composed of three parts: the idiom castles in the air, the semantic orientation a derogatory term representing negative sentiment polarity and a short paraphrase with three general negative words (illusive, impractical and fanciness). The example entry provides us with a labelled sentimental instance with general sentiment features and a negative label. Most importantly, the sentiment polarity of such instance is independent of any context just as the idiom it explains. Based on such observation, another hypothese is proposed as follows:

Hypothesis 2. The sentiment polarity of the idiom paraphrase is independent of domains as the idiom it describes.

With Hypothese 2 admitted, we could constructed a training dataset with the features extracted from paraphrases of idioms as general feature vector, the semantic orientation values as sentimental labels. Then a context-independent classifier could be trained to model the general sentiment knowledge.

4.2 Context Dependent Classifier

As the general sentiment features are only one part of all features in the whole feature space, the other part of context-dependent features should be considered in order to capture the subtle clues embedded in the specific sentiment expressions in tweet context.

To model the context-dependent part of tweet, there are two questions must be solved. The first is to identify the context-dependent part of features extracted from tweet. In fact, new expressions of different opinions appears in microblog with the explosive increase of user-generated content make it a rather difficult problem to clearly tell whether each feature is context dependent or not. However, based on the particular characteristics of tweet, we maked assumption that except for the idioms it contains, sentimental polarity of other words in a tweet be context-dependent in that tweet is limited in 140 words only. The second is

how to find labelled instances to train the context-dependent classifier. Some researchers have proposed distant supervision to solve the training data shortage of Twitter [19, 20], and accordingly in this paper we establish our training dataset by trying to get as many tweet as possible that contains idioms. By stripping off idioms, we extracted left features as context-dependent features, and took the sentiment polarity of idioms as labels. By this way, we got our noisy labelled dataset and a context-dependent classifier was trained to model the context-dependent knowledge.

4.3 Combination of Two Classifiers

Although theoretically the general classifier and context dependent classifier could be able to model different sentiment knowledge separately and classify the sentiment of a tweet accurately to some extent , the coverage and efficiency of such model are limited by the quality and quantity of training resources. Besides, it is obvious that the paraphrase of idiom and tweet segments(lefts by stripping off idioms) is usually short, so the feature vector must be very sparse, which would degrade the effect of such classifiers. For above reasons, a consistent bootstrapping framework of machine learning has been chosen to combine the two general classifiers together. The framework is illustrated in figure 1. As

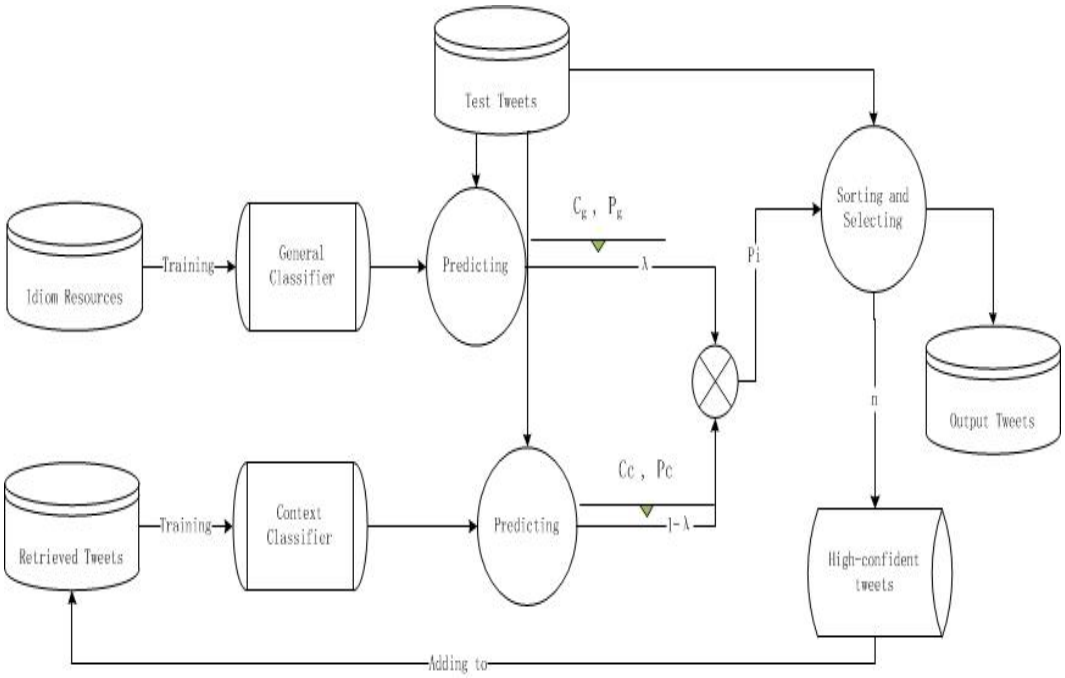

Fig. 1. Combined sentiment classification Framework

could be seen from the framework, a general classifier P_g and a context classifier P_c was applied to test dataset to be classified so that the every test instance x_i were labelled $c_i = \{c_g, c_c\}$ initially and given the confidence $p_i = \{p_g, p_c\}$ by each classifier. Then a combined confidence score was calculated by:

$$p_i = \begin{cases} \lambda * p_g + (1-\lambda) * p_c & \text{if} \quad c_g = c_c; \\ 0 & \text{if} \quad c_g \neq c_c; \end{cases} \tag{3}$$

where λ is coefficient to control weight of different part of features. We initialize $\lambda = 0.5$ with equal weights of general part and context part of features, and to make combined classifier more adaptable for microblog context, we increase the weight of context part step by step with the iteration progressing. The test dataset initially labelled as c_i was sorted by confidence p_i in two sentiment classes $C = \{1, -1\}$ separately. The n positive and negative instances of highest confident score were selected as new training instances to improve the context-dependent classifier to a more context-aware classifier. Such a procedure iterates until convergence. The output of such semi-supervised sentiment classification framework is the predicted results of the sentiment classification. Above all, the whole framework could combine the two classifiers which constructed on divided feature space into a stronger classifier.

5 Experiment

5.1 Experiment Description

Dataset: We crawled from the online idiom dictionary of China Education Network[2] and got an idiom dataset of 8,160 instances labelled with positive and negative sentiment to train the general classifier. From Apr.15th,2013 to May 1st,2013, we monitored Tencent tweet timeline Stream, retrieved and extracted the tweet that contains at least one idiom in our idiom dataset, resulting in about 120,346 instances of tweet. After stripping off idioms from all tweet and removing tweet with words less than 4, a dataset of 91,268 instances was used to train context-dependent classifier. As for test dataset, the dataset of the First Chinese tweet Sentiment Analysis and Semantic Relationship Extraction Evaluation of CCF Natural Language Processing and Chinese Computing[3] was used to evaluate performance of our framwork.

Classifiers and Performance measurement: There are various complicated measurements to evaluate the performance of computational algorithm, of which the simplest accuracy index was chosen to evaluate the performance of our framework, because the comparison between measurements was not the important points of our research. As for classifiers, Naïve Bayes classifier and Maximum Entropy classifier of NLTK (Natural Language ToolKits)[21] package and Support Vector Machine classifier of Libsvm[22] package were used for classification. All the parameters and settings were optimized by cross-validation.

[2] `http://chengyu.teachercn.com`

[3] `http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html`

Baseline and Upper Bound: Two baselines were used to compare with the proposed method, the first one was naïve 50% baseline since the test corpus were balanced with respect to the sentiment classes, the other one was the lexicon-based classifier by comparing positive words and negative words of sentiment lexicon in the same tweet to determine sentiment polarity. As mentioned in section 2, supervised machine learning methods are often setup as upper bound to be challenged by other methods. In the experiments, an upper bound was also setup by training supervised classifiers with the same settings as general classifier except for the dataset settings. Dataset was split five-folded with one fifth for testing and others for training, and the accuracy was calculated by averaging the results of five iterative computations on split dataset.

Preprocessing: Text written in Chinese are not well formatted in that words in a sentence are not separated by space as English. All the text in Chinese must be segmented before bag-of-words features being extracted. In the experiment, Chinese text of train and test dataset was segmented with well-known Chinese segmentation software ICTCLAS[4].

5.2 Result

The results are shown in table 1 in which NB denotes Naïve Bayes classifier, MX denotes Maximum Entropy classifier and SVM denotes Support Vector Machine classifier. From the table the following results can be observed.

Table 1. Results for Different Method

	Lexicon	*Supervised*	*General*	*Context*	*Combined*
NB	0.725	0.785	0.714	0.766	**0.802**
MX	0.725	**0.806**	0.740	0.785	0.802
SVM	0.725	0.826	0.722	0.805	**0.843**

– Firstly, the accuracies of general classifier and context classifier all surpass the nave baseline (50%), which proves that the general classifier is superior to random selection and may be better choice when there are no labelled dataset available for supervised or semi-supervised machine learning sentiment classification.

[4] `http://ictclas.nlpir.org/`

– Secondly, the accuracy of general classifier approximates to the traditional lexicon classifier, because they can both model the general sentiment knowledge. As for context classifier, the performance outperforms lexicon classifier because of its adaption of tweet context.
– Finally, the combined classifier shows the best performance by combining general classifier and context classifer. It even outperforms the upper bound supervised classifier, which proves the effectiveness our proposed framework.

6 Conclusion

Context-dependent problem has always been a main challenge of sentiment analysis. in this paper, we have proposed a novel semi-supervised framework to get it solved in the social media microblog settings. From a different perspective, we carry out the assumption that feature space be divided into the general part and the context part. To make use of two parts of features, two classifiers are trained on dataset constructed from idiom resources and tweets separately. Our framework combines the classifiers with a semi-supervised bootstrapping learning algorithm. The experiment results show that the proposed framework could outperform the state-of-art supervised classifier. In future, we will try to improve the sentiment classification performance by enlarging the context-independent resources and extracting richer features besides bag-of-words feature.

References

1. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2012)
2. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1-2) (2007) 1–135
3. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent Twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 151–160
4. Stieglitz, S., Dang-Xuan, L.: Political communication and influence through microblogging-an empirical analysis of sentiment in twitter messages and retweet behavior. In: HICSS, IEEE Computer Society (2012) 3500–3509
5. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10. EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 79–86
6. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. COLING '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 36–44
7. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: Proceedings of the 22nd international conference on World Wide Web. WWW '13, Republic and Canton of Geneva, Switzerland, International World Wide Web Conferences Steering Committee (2013) 607–618

8. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: Proceedings of the 20th ACM international conference on Information and knowledge management. CIKM '11, New York, NY, USA, ACM (2011) 1031–1040
9. Asiaee T., A., Tepper, M., Banerjee, A., Sapiro, G.: If you are happy and you know it... tweet. In: Proceedings of the 21st ACM international conference on Information and knowledge management. CIKM '12, New York, NY, USA, ACM (2012) 1602–1606
10. Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: Proceedings of the sixth ACM international conference on Web search and data mining. WSDM '13, New York, NY, USA, ACM (2013) 537–546
11. Calais Guerra, P.H., Veloso, A., Meira, Jr., W., Almeida, V.: From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '11, New York, NY, USA, ACM (2011) 150–158
12. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment in short strength detection informal text. J. Am. Soc. Inf. Sci. Technol. **61**(12) (December 2010) 2544–2558
13. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. J. Am. Soc. Inf. Sci. Technol. **63**(1) (January 2012) 163–173
14. Zhao, T., Li, C., Ding, Q., Li, L.: User-sentiment topic model: refining user's topics with sentiment information. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. MDS '12, New York, NY, USA, ACM (2012) 10:1–10:9
15. Baccianella, A.E.S., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (May 2010)
16. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press (1966)
17. Ku, L.W., Liang, Y.T., Chen, H.H.: Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs. (2006)
18. Ku, L.W., Liang, Y.T., Chen, H.H.: Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs. (2006)
19. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision. Technical report, Stanford University
20. Marchetti-Bowick, M., Chambers, N.: Learning for microblogs with distant supervision: Political forecasting with twitter. In Daelemans, W., Lapata, M., Mrquez, L., eds.: EACL, The Association for Computer Linguistics (2012) 603–612
21. Loper, E., Bird, S.: NLTK: The Natural Language Toolkit (May 2002)
22. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27 Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.