DOI: 10.7763/IPEDR. 2013. V68. 1

Sentiment Classification in Chinese Microblogs: Lexicon-based and Learning-based Approaches

Bo Yuan, Ying Liu and Hui Li[+]

Laboratory of Computational Linguistics, Tsinghua University, Beijing, China

Abstract. Sentiment classification in Chinese microblogs is more challenging than that of Twitter for numerous reasons. In this paper, two kinds of approaches are proposed to classify opinionated Chinese-microblog posts: 1) lexicon-based approaches combining Simple Sentiment Word-Count Method with 3 Chinese sentiment lexicons, 2) machine learning models with multiple features. According to our experiment, lexicon-based approaches can yield relatively fine results and machine learning classifiers outperform both the majority baseline and lexicon-based approaches. Among all the machine learning-based approaches, Random Forests works best and the results are satisfactory.

Keywords: Sentiment Classification, Chinese microblog, Sentiment Lexicon, Machine Learning

1. Introduction

Sentiment classification refers to the automatic categorization of texts into 3 polarity classes, i.e. “positive”, “negative” or “neutral”. With more and more opinionated texts available on the World Wide Web, the mining of sentiment, emotion, comments, attitudes etc. becomes alluring due to its potential intellectual and economic value. Sentiment analysis therefore becomes an important branch of Natural Language Processing. However, since microblog is inherently limited in length, the information a post can provide is much too insufficient. Thus microblog poses new challenges to Sentiment Analysis [1]-[2].

In China, microblog is generally referred to as “Weibo”, which will be used in our paper with the term "microblog" interchangeably. The research of Twitter Sentiment Classification is relatively more sophisticated and comprehensive. Weibo Sentiment analysis begins later and is even more challenging concerning the characteristics of Chinese language and the website. First, although both Weibo and Twitter have a 140-character limit, most Chinese words consists of only 2 or 3 characters with no space between them. Therefore a Weibo post can convey far more information than is possible in Twitter. Another problem we should concern is register [3] diversity in social networking websites. The current situation of Chinese Language Processing also makes it hard to extract accurate features, which will be discussed in section 3.

The rest of this paper is organized as follows: Section 2 reviews briefly some related work. In Section 3 and 4, lexicon-based and machine learning-based approaches are introduced in detail. Experiments and results are reported in Section 5. Finally, Section 6 gives the conclusion and an outlook for our future work.

2. Related Work

Sentiment Analysis and Opinion Mining refers to the process of mining, extracting, classifying, retrieving or visualizing the views, emotions, comments, attitudes etc. in human language [4]. The study of sentiment classification is one of its subtopic with the longest history. Sentiment classification is a general term for both subjectivity and polarity detection, wherein the input texts are first tagged with “subjective” or “objective” to denote its subjectivity, among which subjective texts are classified as “positive” or “negative”. Sentiment classification ranges from sentence-level to text-level. Rule-based approaches are applied where lexicon-based methods [5] and pattern-based methods [6] are popular in earlier stages. Machine learning models are more common in this task. Supervised learning approaches [7], unsupervised approached or semi-supervised approaches [8] have been applied in different corpus domains.

Earlier microblog sentiment classification work mainly utilizes learning-based classifiers and text-categorization features. Later work began to take other features beyond bag-of-words into consideration,

[+] Corresponding author: *E-mail address*: yuanbo@outlook.com.

besides linguistic features, abstract representation has been leveraged, such as the meta-information of words, the characteristics of Twitter [9]. Recently target-dependency [10] and context awareness [11] has caught much attention. However, there is very little literature on Weibo sentiment classification.

Sentiment Lexicon plays a vital role in sentiment analysis tasks. Linguistic Inquiry and Word Count (LIWC) [12], General Inquirer Lexicon[13] are two of the most widely-used sentiment lexicon and SentiWordNet[1] combines polarity information with semantic information of English words. In Chinese, we have Hownet Sentiment Dictionary[2], National Taiwan University Sentiment Dictionary (NTUSD)[3] and Chinese Emotion Word Ontology (CEWO) [14].

3. Lexicon-Based Approaches

3.1. Simple Sentiment Word Count Method

Simple Sentiment Word-Count Method (SSWCM) is an intuitively basic algorithm for sentiment classification. Polarity of an input text is determined by the number difference of sentiment words, as shown in Table 1 where #pos and #neg each represents number of sentiment words in the specific text.

Table 1: Simple sentiment word count method

Polarity	*Condition*
positive	#pos > #neg
negative	#pos < #neg
neutral	#pos = #neg

The performance of SSWCM relies on 1) a highly-domain-suited sentiment lexicon and 2) a well-designed Chinese word-segmentation system, both of which have a long way to go.

3.2. Sentiment Lexicons

The composition of sentiment lexicons we use in our experiment is shown as in Table 2: 1).The original HSD consists of four lists of sentiment words. In our experiment, the positive/negative emotion and opinion lexicon are merged into a positive/negative sentiment lexicon and the overlap part is removed. 2). As for NTUSD, we convert the words and phrases from traditional Chinese to simplified Chinese manually. 3). CEWO was created on the basis of 7 emotion classes with 21 sub-classes. We combine lists of words according to their semantic orientation. In our experiment, "objective" words which don't convey certain emotion and "neutral" words which can be used both as positive are not taken into account.

Table 2 Sentiment Lexicons in Experiments

Sentiment Lexicon Name	Description
Hownet Sentiment Dictionary (HSD)	4528 positive entries, 4320 negative entries)
National Taiwan University Sentiment Dictionary (NTUSD)	2810 positive entries, 8276 negative entries
Chinese Emotion Word Ontology (CEWO)	11229 positive entries, 10784 negative entries

All the sentiment lexicons incorporate most commonly used emoticons extracted from 1 million Weibo posts, among which 103 are positive and 73 are negative.

4. Machine Learning-Based Approaches

4.1. Feature Selection

4.1.1. Bag-of-Words and Sentiment Lexicon Features

In this part, we segment and POS tag those posts and merely selected those *content words*. This is due to the fact that most "concrete" meaning is conveyed by content words rather than function words. Sentiment

[1] http://sentiwordnet.isti.cnr.it/
[2] http://keenage.com/
[3] http://nlg18.csie.*ntu*.edu.tw:8080/opinion/index.html

lexicon entries are also included in the word list before the conversion. Bag-of-Words feature terms and Sentiment Lexicon Features are weighted by IDF and pruned by default method of Weka software.

4.1.2. Multiple Features

Table 3 is divided into 3 parts. The first part is linguistic features. We manually extract sentiment-related syntactic patterns and collect 165 Chinese adverbs of degree, 37 negative adverbs. In addition, regular expressions are used to extract some specific punctuation usage by Weibo users. The second part contains Weibo characteristics. The third part lists context features because sentiment analysis over user-generated data in social websites is more context-dependent compared to traditional NLP tasks. The terminology "context" is by and large a pragmatic one where more and more features need to be extracted. Far more, work needs to be done concerning the context feature mining on sentiment analysis.

Table 3: Multiple features in Weibo Sentiment Classification

FEATURE NAME	VALUE SPACE	DESCRIPTION
_adv	(0,1)	If the post contains adverbial phrases of degree
_neg	(0,1)	If the post contains negative adverbial phrases
_punc	(0,1)	If the user uses question mark, exclamation, or more than one punctuation mark at one time such as "!!!!!!!!", "??????" or "。。。"
_ hashtag	(0,1)	If the post contains hashtag or not. Unlike Twitter hashtag, Weibo hashtag is formed by at least two "#", which can be an indicator of emotion and topic.
_link	(0,1)	If the post contains a link.
_ mention	(0,1)	If the post mentions (@) another user or not.
_ Sharedcontent,	(0,1)	If the user shared a picture, audio or video.
_ repost	(0,1)	If the post is replied by other users.
_reply	(0,1)	If other users reposted the message.
_like	(0,1)	If the post is "liked" by other users by clicking the button under it.

4.2. Machine Learning Models

We trained Naïve Bayes Classifier, Maximum Entropy Classifier and Random Forests Classifier incorporating the features selected in the previous section.

Naïve Bayes Classifier is the simplest Bayesian classier in various machine learning tasks. Although popularly applied to many tasks, the generative model is frequently questioned for its assumption that all attributes of the examples are independent of each other given. Maximum Entropy Models (MaxEnt) does not assume features are independent. Discriminative models like MaxEnt take the data as given, and put a probability over hidden structure given the data. Another model we have included in our experiment is Random Forests, a machine learning ensemble consisting of a bagging of unpruned decision trees.

The software we have implemented our algorithms with are Weka and Stanford Classifier[4], and in all experiment the default setting is used.

5. Experiment

5.1. Dataset and Annotation

Table 4: Basic Information about

	Positive	Neutral	Negative	Total	Remarks
DSD-1	579	1821	821	3221	IT-Products
DSD-2	473	1651	738	2862	Movie
DSD-3	447	1313	534	2294	Fast Consumption Product
DSD-4	532	1572	1124	3228	Celebrity
RD	938	1211	885	3034	Random Selected Posts

We design a dataset consisting of 4 Domain Specific Datasets (DSD) and 1 Random Dataset (RD). Those datasets have been labelled with polarity for 4 weeks in a row by 2 annotators, who are both graduate

[4] http://nlp.stanford.edu/software/classifier.shtml

school students and native Chinese speakers. Before annotation, spams are manually filtered and a training seminar is held. 200 examples is annotated by both when 17.0% disagreement is found and all of them are "neutral vs. not-neutral" differences. No "positive vs. negative" disagreement is reported.

5.2. Approaches and Evaluation

Various methods are used for sentiment classification, including lexicon-based approaches and machine learning-based approaches. We use "Majority-Baseline" where all examples are labelled with the most labels in the datasets. For lexicon-based approaches, Simple Sentiment Word-Count Method (SSWCM) is applied with three different sentiment lexicons. For learning-based approaches, the features in section 4 are extracted and 3 distinct models are trained.

We use *Accuracy* to evaluate the classifiers. Accuracy represents the total percentage of right examples classified out of all examples. For learning-based methods, 6-fold cross-validation is conducted and the final results are based on their average accuracy.

5.3. Results and Discussions

As depicted in Table 5, the overall accuracy of lexicon-based approaches is beyond our expectation. Since data distribution is imbalanced where the majority of examples are neutral, the baseline result is high in accuracy. In DSD-1 to 3, the accuracy of our baseline are all above 0.5. *SSWCM-HSD* exceed the baseline 3 times in DSD-2/3 and RD. SSWCM-NTUSD achieves twice over 0.6000 in accuracy. The best result of lexicon-based approaches is 0.6343, achieved by SSWCM-CEWO in DSD-3, though one shot given by SSWCM-CEWO is below baseline (in DSD-1). All the lowest accuracies of lexicon-based classifiers are in RD dataset where all results are below 0.5000.

Table 5: Experiment Results of Classifiers

	DSD-1	DSD-2	DSD-3	DSD-4	RD
Baseline	*0.5654*	*0.5654*	*0.5724*	*0.4870*	*0.3991*
SSWCM-HSD	0.5352	0.5834	0.5767	0.4634	0.4684
SSWCM-NTUSD	0.5964	0.6076	0.6160	0.4957	0.4285
SSWCM-CEWO	0.5560	0.5843	**0.6343**	0.5505	0.4604
Naïve Bayes	0.5973	0.5874	0.5628	0.5489	0.4535
MaxEnt	0.6243	0.6426	0.6146	0.5666	0.5069
Random Forests	0.6541	**0.6796**	0.6600	0.5917	0.5267

Basically all the learning based classifiers work better than the baseline, except Naïve Bayes classifier in DSD-3. Across all the datasets, Random Forests Classifier achieves the best accuracy and far outperforms the baseline. In DSD-2, Random Forests Model works best in the whole experiments (0.6796).

A comparison of the two methodologies are drawn in Figure 1, where LEX/LEA-Average refers to the average accuracy of lexicon-based or learning-based approaches; and LEX-Max and LEA-Max refers to the maximum accuracy achieved either of the approaches. Almost all approaches across 5 datasets achieve average accuracies that is higher than the baseline. The only case that fails is lexicon-based approach in DSD-1 datasets. All maximum accuracies exceed the baseline and ranges above 0.5000. In general, Domain-Specific Datasets sees better performance than Random Datasets. Machine learning approaches outperform lexicon-based approaches by around 0.1000 in average. But when it comes to the maximum accuracy, the difference of this two kinds of approaches are not significant, although in RD dataset the difference is around 0.5000. A higher performance can be anticipated if parameters of models are better adjusted.

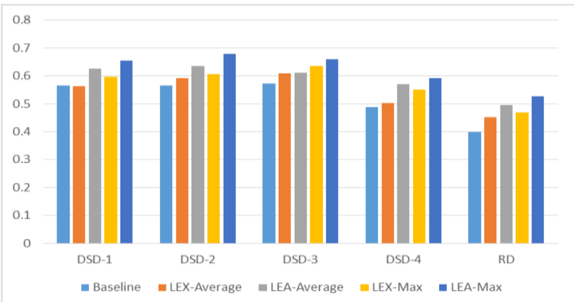

Fig. 1: Comparison of Lexicon-based and Learning-based Approaches

6. Conclusion

In this paper, we initiate a variety of approaches in the task of sentiment classification over Weibo. We dive into lexicon-based approaches by combining SSWCM with three Chinese sentiment lexicons. Meanwhile 3 state-of-the-art machine learning models are applied in sentiment classification. An overall comparison among different approaches is drawn. According to our experiment, SSWCM can provide a relatively fine result and performs better than the baseline. Machine learning classifiers with multiple features work better and outperform lexicon-based approaches. We also find that generally all approaches work better in Domain-Specific Datasets than in Random Datasets.

Our future work will be the optimization of machine-learning models and combination of both approaches. The need for domain-adaptable sentiment lexicons and more sophisticated feature extraction drives us further in our research. The last but not least, more efforts should be made into the foundation building of Chinese Language Processing.

7. Acknowledge

The work is supported by the project, Model Research on Interactive Behavior and Linguistic Features Based on Pragmatic Information, the National Natural Science Foundation of China, 2012-2015 (61171114)

8. References

[1] Xie L X. Sentiment analysis of Chinese micro blog using SVM[D]. Master Thesis. Tsinghua University, Beijing, 2011.

[2] Bermingham A, Smeaton A F. Classifying sentiment in microblogs: is brevity an advantage?[C]//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 1833-1836.

[3] Leckie-Tarry H. Language and context: a functional linguistic theory of register[M]. Pinter Pub Ltd, 1995.

[4] Liu B. Sentiment analysis and opinion mining[J]. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1-167.

[5] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.

[6] Riloff E, Wiebe J. Learning extraction patterns for subjective expressions[C]//Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 2003: 105-112.

[7] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[8] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 417-424.

[9] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 36-44.

[10] Jiang L, Yu M, Zhou M, et al. Target-dependent Twitter Sentiment Classification[C]//ACL. 2011: 151-160.

[11] Aisopos F, Papadakis G, Tserpes K, et al. Content vs. context for sentiment analysis: a comparative analysis over microblogs[C]//Proceedings of the 23rd ACM conference on Hypertext and social media. ACM, 2012: 187-196.

[12] Pennebaker J W, Francis M E, Booth R J. Linguistic inquiry and word count: LIWC 2001[J]. Mahway: Lawrence Erlbaum Associates, 2001: 71.

[13] Stone P, Dunphy D C, Smith M S, et al. The general inquirer: A computer approach to content analysis[J]. Journal of Regional Science, 1968, 8(1).

[14] Yan J, Bracewell D B, Ren F, et al. The Creation of a Chinese Emotion Ontology Based on HowNet[J]. Engineering Letters, 2008, 16(1): 166-171.