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 labeled idiom resources, and a context classifier on context-dependent part of features is trained on random tweets retrieved from microblog in distant supervision manner. Then a semi-supervised framework is developed to combine the general classifier and context classifier into a bootstraping classifier. Experiments results show that the proposed framework is effective and achieves encouraging accuracy performance.

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

1 Introduction

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

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 network platforms 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 freely. It has been well recognized that user-generated content of microblog with rich sentiment information can trigger more attention, feedback or participation and have attracted notice of sentiment analysis researchers[**?**]. 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 or 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 online communications for users, it is difficult for computers to accurately identify the semantic meanings of these words. To make situation worse, new words may arise and old words may change their meaning continously 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 the text of microblog always depends on such particular expressions as emoticons, repeated letters and exclamation, etc. So there are always no universal labeled resources available for sentiment classification. But acquiring human-labeled data is costly and difficult, for manual annotation is very expensive and time-consuming.

Context dependence has been the main challenge Sentiment classification researches of tweets face. In this work, we have focused on the context-dependent problem of sentiment classification in the settings of microblog. To solve the problem, we propose a novel semi-supervised framework based on our two rational hypotheses. By dividing the feature space into two parts, a general classifier on general part of features is trained on off-the-shelf labeled 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 bootstraping classifier. The experiments carried out in a Chinese microblog platform show that the framework is effctive and achieves encouraging accuracy performance.

The rest of the paper is organized as follows: Related works are discussed in section 2. Our two hypotheses for the feature space 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 hot research area for years. Previous research mainly focuses on product or movie reviews. Generally, in terms of methodology, rule-based approaches and machine learning 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[**?**,**?**].
Recently, the sentiment analysis research has begun to pay more and more attention to the massive user-generated content of social networks such as Twitter[1]. Barbosa and Feng[**?**] first investigated to use a two-stage SVM classifier which proved to be more robust regarding biased and noisy data. Many studies show that some unique characteristics of Twitter can also be incoporated into sentiment analysis. Hu et al.[**?**] 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.[**?**] 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.[**?**] 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. Their investigation illustrated that three types of information is useful to address the task including sentiment polarity of tweets containing the hashtag, hashtags co-occurrence relationship and the literalmeaning of hashtags. Amir Asiaee T et al.[**?**] 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.[**?**] 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 that argue that humans tend to have consistently biased opinions, Guerra et al.[**?**] 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.[**?**,**?**] designed SentiStrength, an algorithm for extracting sentiment strength from informal English text. The algorithm exploits the grammar and spelling styles in typical microblogs and builds on human-evaluated dictionaries for words connotated with positive or negative sentiments. Zhao et al.[**?**] 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 Twitter context by making use of the network and language characteristics of moricroblog, no matter what arroaches they have taken. However, in this paper, we solve the context-dependent problem from a novel perspective by seperating context-independent part of features from context-dependent part.

[1] `twitter.com`

3 Problem Formulaiton and Solution

3.1 Problem Formulation

Simply speaking, sentiment classification aims to classify text as predefined sentiment binary 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 $x = R^n$, where n is the size of a pre-specified feature volumn V. In sentiment classification, the weight of each entry in the vector usually is often specified as binary, with weight equals to 1 for terms 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, a tweet from famoous English microblog platform Twitter[2] is listed as following:

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

If simple features are considered, all words would be extracted as equal features to model the sentimental polarity when computationally classifying sentiment polarity of this tweet. However with careful considerations, we could find that such words as "@Kid_Cloudz, :D, lovee, everrrrr,!!!!!" are more frequently used in the microblog context, while "Happy, birthday, wish, best, thanks" are negative sentiment indicators of text almost in any context. With such intuition, we proposed an feature space division hypothese as:

Hypothese 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 polarity independent of any context;*
- *context-dependent part, i.e. context features, which indicate sentiment polarity of text in specific context.*

[2] `twitter.com`

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

$$x = \{ \begin{matrix} x_g & : \text{generalfeatures} \\ x_c & : \text{contextfeatures} \end{matrix} \tag{2}$$

where $x_g = (w_1, \ldots, w_l)$ denotes the general part of feature space; $x_c = (w_{l+1}, \ldots, w_v)$ denotes the context dependent part.
Based on hypothese 1, we studied following questions in this paper:

- how to identify each part of feature space, and what's the meaning to the sentiment classification of tweet?
- how to model sentiment polarity of tweet with the two parts of features in the machine learning framework?
- how to find seperate training instances for sentiment classifiers based on the two parts of features?

4 Sentiment Classification Framework

Besides normal expressions, some languages of microblog are often considered "Mars Language" because of their obscurity. But we could sometimes distinguish the sentiment polarity of the tweet even if some words seem strange to us. Intuitively, this kind of phenomenon may be explained by the general part of the text which is used to express the holistic sentiment polarity of the author in any context, and the polarity of general sentiment words are prone to be recognized by anyone independent of context knowledge. Comparably, in sentiment classification, we put forward that, the sentiment polarity of a document could sometimes still be recognized with only general part of feature spacex_g. 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[**?**] and General Inquiry[**?**] in English, Hownet[**?**] and NTUSD (Chinese Network Sentiment Dictionary) in Chinese[**?**], 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 labeled 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:

Hypothese 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 labeled instances to train the context-dependent classifier. Some researchers have proposed distant supervision to solve the training data shortage of Twitter [**?**,**?**], 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 labeled 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 seperately 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

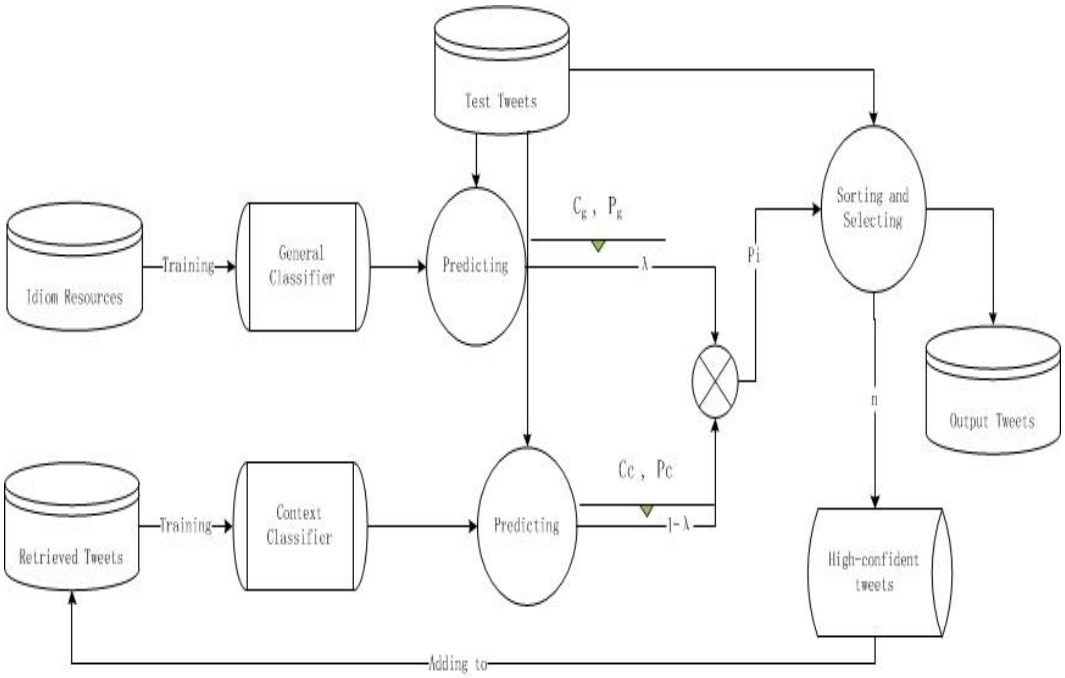

Fig. 1. Combined sentiment classification Framework

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 labeled $c_i = \{c_g, c_c\}$ 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 labeled as c_i was sorted by confidence p_i in two sentiment classes $C = \{1, -1\}$ seperately. 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[3] and got an idiom dataset of 8,160 instances labeled 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[4] 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)[**?**] package and Support Vector Machine classifier of Libsvm[**?**] package were used for classification. All the parameters and settings were optimized by cross-validation.

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

[4] `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[5].

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 labeled dataset available for supervised or semi-supervised machine learning sentiment classification.

[5] `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 seperately. Our framework combines the classifiers with a semi-supervised bootstraping 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.