

Topic-Based Chinese Message Sentiment Analysis: A Multilayered Analysis System

Hongjie Li Zhongqian Sun Wei Yang

Tencent Intelligent Computing and Search Lab

Nanshan District, Shenzhen P.R. China

{hongjieli, sallensun, willyang}@tencent.com

Abstract

Sentiment analysis in social media has attracted significant attention. Although researchers have proposed many methods, a single method is hard to meet requirement in industrial applications. In this paper, based on massive data of Tencent and industrial practice, we present a multilayered analysis system (MAS) on social media. The system is composed of three sub-systems, including topic correlation calculation, topic-related sentence recognition and sentence polarity classification. Each sub-system is composed of several simple models. Also, we have set up a closed-loop feature mining and model updating system, which will continuously promote performance of MAS. In addition, this offline system requires very little intervention. The system, including online and offline parts, has been applied in several practical projects and obtained the best results in the evaluation of task 2 of SIGHAN-8.

1 Introduction

The popularity of Web 2.0 applications promotes the emergence of user generated content (UGC), e.g., the comments in blogosphere, and the UGC reflects the viewpoints of web users towards a specific event or product. Scholars have carried out a series of studies around these data, especially in the research of sentiment analysis. It aims to understand the subjective opinions of characters, events and other subjects based on the analysis of the content published by users. Sentiment analysis has a wide range of applications, e.g. the social public opinions, the word of mouth analysis, potential users mining.

In this article, we focus on sentiment analysis of short-text generated by users, for example, micro

blog, news comment, products comment, tweets and so on. Many researchers have proposed many methods to improve the effect of sentiment analysis. Mei (2007) introduced latent semantic analysis model for sentiment analysis, e.g. LDA. Si et al. first utilize a continuous Dirichlet Process Mixture model to classify tweets. A supervised sentiment classification framework was proposed by (Davidov et al., 2010). Based on KNN, they use Emoticons and hashtag to classify sentiment in tweets. Another significant effort for sentiment analysis is proposed by (Barbosa and Feng, 2010) who use polarity predictions from three websites as noisy labels to train a SVM model. Hassan (2010) use dependency relations and part-of-speech patterns to classify the message in Usenet with Supervised Markoff model. A.Meena (2007) analyzes the impact of conjunction to the emotion analysis, but the system do not have field adaptive ability. Socher (2011) extended word representations beyond simple vectors. They merge words in sentences and create phrase representations recursively.

In industry, a single model is hard to achieve the expected performance. Based on massive data of Tencent, we propose an multilayered approach which integrates multiple simple methods. Meanwhile, we set up a closed-loop offline mining process, which optimizes the online classification results through continuously mining new features. The approach has been tested in task 2 of SIGHAN-8. And the result showed that both the precise and recall improved a lot.

2 A Multilayered Anasysis System

In this section, we first introduce online part of MAS. Then we introduce how MAS forms a closed-loop updating system. Last, we present some key points of MAS.

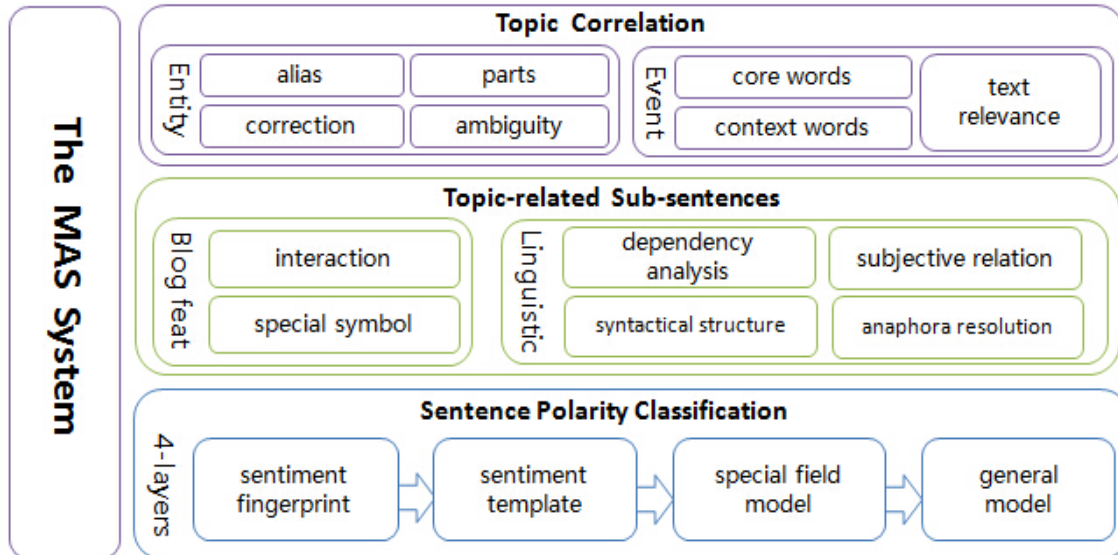


Figure 1: The online methods of MAS.

2.1 The Online Methods of MAS

As shown in Figure 1, MAS is composed of three sub-systems, including topic correlation calculation, topic-related sentence recognition and sentence polarity classification.

2.1.1 Topic Correlation Calculation

This system is used to decide whether a message is associated with the specific topic. Here, we divide topics into two types. One type is “Entity”, such as “三星S6”, “苹果手机” etc. The other type is “Event”, such as “香港占中”, “中国人疯抢日本马桶盖” etc. Different approaches are used to process the two type topics.

Entity Topic Correlation. The existing of entity name in messages determines the entity topic correlation. The difficulties of this problem is alias or varietas recognition, and word sense disambiguation. For example, topic “三星Galaxy S6” is usually expressed as “三星S6”, “Galaxy S6”, “samsung s6” and “三桑S6” etc. A message containing any of the expressions is regarded as a correlated message. But the simple approach is embarrassed when the entity is ambiguous. Take the topic “苹果手机” into account, “苹果” in Chinese may refer to a kind of fruit, or refer to apple cellphone. Therefore, we need eliminate this ambiguity.

We use the context of entity to solve the ambiguity. In simple terms, if “苹果” appears with “吃了”, then it’s more likely referring to fruits. If it is co-occurred with “屏幕”, then it’s more likely to be cellphone. Formally, suppose $D = \{d_1, \dots, d_k, \dots, d_m\}$ is a sentence and d_i

denotes the i th word of sentence. And d_k is the specific entity. Sentence will be divided into two parts, $\{d_1, \dots, d_{k-1}\}$ and $\{d_{k+1}, \dots, d_m\}$. We count words appear in the two parts separately, as well as the co-occurrence of words each of which comes from the two different parts. Firstly, they are counted in a labeled dataset. Then, we statistic them in a bigger data set. Finally, using TF-IDF method, we select features as the topic’s context.

We call the entity recognition and correction API of Tencent Wenzhi to solve alias or varietas problem.

Event Topic Correlation. For event type topic, we first extract the core words of events. Then, extend the words to many context words and phrase which has close relationship with event. Finally, text correlation algorithm is used to calculate the event topic correlation.

2.1.2 Topic-Related Sentence Recognition

This strategy is used to recognize sub-sentences that relate to special topic from message and get rid off non-related sub-sentences.

In this evaluation task, two kind of approaches are included. One kind is special characters of Micro-blog (e.g. replying relation), and the other approach relies on NLP technologies, such as subjective relation extracting, dependency parsing, sentence analysis (such as comparative sentence, interrogative sentence etc.) and so on.

2.1.3 Sentence Polarity Classification

In this section, we propose a 4-layer classification system. It gives the polarity of a sentence, positive, negative or neutral. Meanwhile, for each

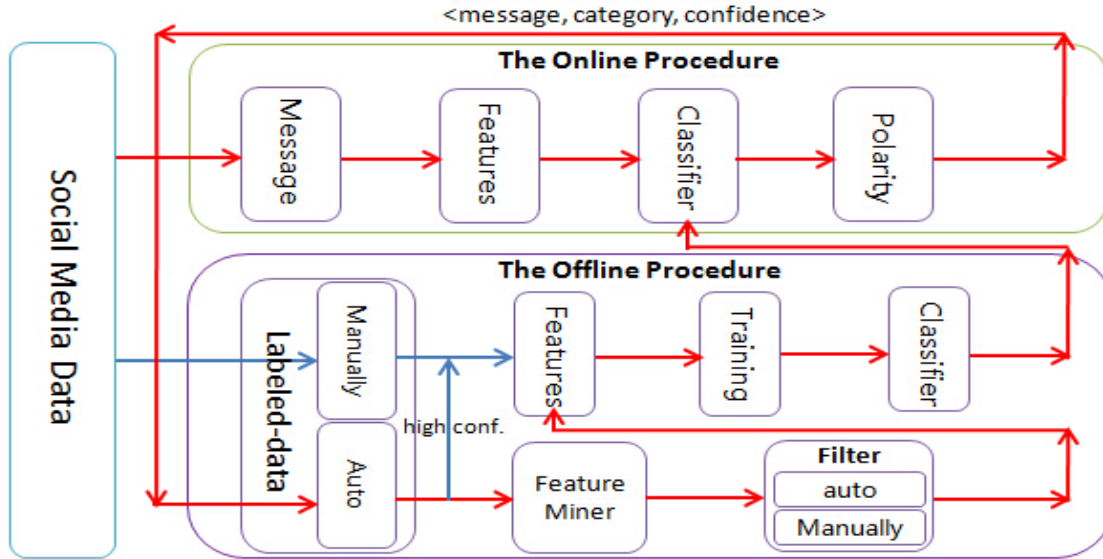


Figure 2: The online methods of MAS.

layer, it is composed of online and offline system. The offline system will continuously mine new features and updating models to promote the online system’s performance.

The four layers of classification system are sentiment fingerprint layer, sentence template layer, specific field model layer and general model layer.

Sentiment Fingerprint Layer. It aims to mine the idioms and popular expressions, e.g. “人在做，天在看！”，“我也是醉了”，“人生如戏，全靠演技”。 These expressions are usually hard to extract valid features for classification. In our approach, we firstly mine these expressions offline, manually label their polarities and generate a sentiment fingerprint database. When we classify a sentence, it will be looked up in our fingerprint database first.

Sentence Template Layer. It focus on lexical collocation when people express their emotions, e.g. “希望...去死”，“希望...坚强”，“没...那么差”。 These lexical collocation jointly reflect people’s emotion. If they are separated into single words, the sentiment may totally different. For example, “希望在苦难中挣扎的人们坚强” is positive emotion, but it is easily identified as negative due to the words “苦难” and “挣扎”。 And the sentence templates can avoid it.

Special Field Model Layer. It’s used to classify messages from specific field, such as movie, music, app and so on. It will use more specific features than general model. For example, in app field, “闪退” and “卡顿” are negative appraisal for app’s stability. These words are very strong features for special field model. But they

have little sense in general model. Therefore, in specific field, special field model usually get better performance than general model, because it can model more domain knowledge. We will present the details together with general model, as they using similar algorithms and features.

General Model Layer. It will classify messages that previous layers can’t handle. It has the highest recall rate and lowest accuracy than the first three layers. It is composed of multiple algorithms and kinds of features. Formally, $g(x) = \sum_i \alpha_i f_i(x)$, where f_i is the i th model and α_i is the weight of f_i . The models are selected from a basic algorithm pool and the pool contains several different approaches, including Bayesian, SVM, Neural networks. And the weight are trained based on the training data.

2.2 The Closed-loop Updating System

In order to continuously promote the performance of MAS, we have set up a closed-loop feature mining and model updating system. And this offline system requires very little intervention. The general method is shown in Figure 2.

As shown in Figure 2, the online system processes messages from different projects and label them with confidence. Then the processed messages are sent to offline system as training data. The offline system divides data into two sets. One set contains the high confidence messages and the other contains lower confidence messages. The high confidence messages are merged to training data directly. The low confidence messages are send to human. Then they are labeled and merged

to training data. The offline system process these data to mine new features and update models. It's worth mentioning that new features are firstly directly add to models without manual confirmation. And we verify new model with test data, if both recall rate and accuracy are better, the online model will be replaced by new model. Otherwise, we will manually analyse and update model.

2.3 Some Key Points of MAS

In this section, we will introduce some keypoints of MAS. Firstly, we will present algorithm used in system. Then, features in MAS are introduced. Finally, we show that how features are mined.

2.3.1 Algorithms

Naive Bayesian. Naive Bayesian is the simplest but effective classifier. Here, we use sentiment phrase as features. Because each category can be considered having the same prior probability $P(c)$, the probability of a phrase d in category c can be expressed as likelihood $p(d|c)$. Based on independence assumption, the probability of a sentence D belonging to c can be calculated by $p(c|D) = \prod_{d \in D} p(d|c)$.

SVM. Svm is an effective classifier which can achieve good performance in high-dimensional feature space. The samples are represented as a point in space, and SVM divides these samples by a clear margin as wide as possible. In this work, `libsvm[rf]` are used to train a classifier. The option of probability estimation in `libsvm` is turned on, therefore it can produce a probability of class c given a sentence x , i.e. $P(c|x)$. For each sentence, we take N-Gram features and PMI lexicons as features.

Neural Network. Neural network is a nonlinear statistical data model. It can effectively model the relation between input and output. And it's one of the most often used algorithm for classification. In this work, we use the open source tool FANN to train a classifier. The classifier using the same features as SVM classifier.

2.3.2 Features

Word N-gram: We select N-gram (bigram and trigram) features from messages using feature selection algorithm, such as TF-IDF, χ^2 -Test and so on. When a certain N-gram appears in the message, the corresponding feature is set to 1, otherwise 0. The training data scale is 1.5 million and finally select 500 thousand features.

PMI bigram lexicons: Some lexicons often appear in sentences together. They determine the polarity of sentence jointly and one lexicon may express different emotion, sometimes even opposite. The features are generated base on point-wise mutual information(PMI). Then, we choose the most relevant features using the same approach with **Word N-gram**. We finally choose 50 thousand features for each category.

Sentiment Phrase: It has been shown that words with positive or negative emotions are important to sentiment classification. In this work, we believe that phrases with emotions are more useful, and we simply extend words (e.g. 喜欢) to phrase (e.g. 喜欢的不得了). Based on Tencent massive data, the words and phrases are mined automatically. Up to now, there are more than 70 thousand phrases collected.

3 Experiments

We used the dataset provided by task 2 of SIGHAN-8 to evaluate our model. The dataset contains about twenty thousand weibo comments of 20 topics. According to the official standard set, we tested performance on each topic, and the results are shown in Table 1 and Table 2.

It is shown in Table 1 that the overall performance of MAS on all given topics and the median of performance of all teams. The $F1$ value of positive and negative emotion of MAS are 60.39% and 69.38%, which are significant better than 19.15% and 36.46% of median value.

In Table 2, the best 3 and worst 3 performance of topics are chosen. The best topics, with $F1$ value around 70%, are all Entity-Topics, e.g. “12306验证码”, “陶华碧” and “美图手机”. Meanwhile, the worst topics, whose average $F1$ value is around 20%, are regard as Event-Topics, e.g. “中国政府也门撤侨”, “隆平高科超级稻” and “就业季”. In the worst cases, some Event-Topics are even classified with none message correct. For example, the $F1$ value of negative sentiment of “中国政府也门撤侨” and positive sentiment of “隆平高科超级稻” are both 0. Therefore, the MAS can deal with Entity-Topics better than Event-Topics. The main reason leading to the result is that it is difficult to determine whether a message is related to the Event-Topic then Entity-Topic. And it's an aspect that should be improved further.

| | Positive | | | Negative | | |
|--------|-----------|--------|--------|-----------|--------|--------|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| MAS | 58.80% | 62.07% | 60.39% | 79.17% | 61.75% | 69.38% |
| Median | 19.97% | 23.37% | 19.15% | 44.00% | 34.56% | 36.46% |

Table 1: The overall performance of our method and the median of all teams.

| Topic | Positive | | | Negative | | |
|----------|-----------|--------|--------|-----------|--------|--------|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| 12306验证码 | 62.12% | 87.23% | 72.57% | 94.13% | 78.34% | 85.51% |
| 陶华碧 | 91.20% | 59.07% | 71.70% | 82.05% | 86.49% | 84.21% |
| 美图手机 | 88.24% | 54.97% | 67.74% | 84.21% | 57.14% | 68.09% |
| 就业季 | 16.67% | 25.00% | 20.00% | 85.51% | 30.10% | 44.53% |
| 中国政府也门撤侨 | 63.64% | 50.00% | 56.00% | 0.00% | 0.00% | 0.00% |
| 隆平高科超级稻 | 0.00% | 0.00% | 0.00% | 28.36% | 43.18% | 34.23% |

Table 2: The best 3 and worst 3 performance of MAS on topics of SIGHAN-8.

4 Conclusion

In this paper, we propose a multilayered analysis approach, which is proved to be effect for sentiment analysis. In our method, the online and of-line procedures are formed a closed-loop system to continuously improve approach’s performance. And this system can be easily used to any classification work. However, the correlation between topic and message is still a limitation, especially between Event-topics and messages. It is one of the most important optimization in our further work.

References

- Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics.
- Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics.
- Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. 2010. What’s with the attitude?: identifying sentences with attitude in online discussions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1245–1255. Association for Computational Linguistics.
- Arun Meena and TV Prabhakar. 2007. *Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis*. Springer.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics.