Label Micro-blog Topics Using the Bayesian Inference Method

Heng Gao[1], Qiudan Li[1], and Xiaolong Zheng[1,2]

[1] The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences,
Beijing, 100190, P.R. China
[2] Dongguan Research Institute of CASIA, Cloud Computing Center,
Chinese Academy of Sciences, Songshan Lake,
Dongguan, 523808, P.R. China
{heng.gao,qiudan.li,xiaolong.zheng}@ia.ac.cn

Abstract. Classifying Micro-blog content is a popular research topic in social media, which can help users access their favorite information quickly. Much research focuses on classifying Micro-blog content with short text dataset. The challenge is the classification effect may be hampered by content ambiguity. To address this challenge, we propose a novel classification framework using external knowledge base and the Bayesian inference method. We first introduce Baidu Encyclopedia to extract effective features, and then we train efficient classifiers with the probabilistic graphic model. The proposed model can classify micro-blog content reliably. Experiments on Sina-weibo dataset demonstrate the effectiveness of the proposed method.

Keywords: Micro-blog, classification, Bayesian inference, information organization.

1 Introduction

Nowadays, popular social media sites including Facebook, Twitter, YouTube and Flickr are attracting more and more people to participate in [1, 2, 3]. In China, Sina-weibo, an important Micro-blog platform, is also growing constantly. As of December 2012, Sina-weibo owns 500 million active users, and these users publish over 100 million tweets every day[1]. With users' active participation, Sina-weibo becomes one of the most important information sharing platforms in China. As Sina-weibo users become very passionate with online information acquiring, they often find it hard to acquire their favorite content. On one hand, they have to browse through a lot of posts to find their interested ones. On the other hand, the mixed information flow with various types may distract their attentions.

In such circumstances, classifying Micro-blog content becomes a necessary approach for users' information demand. Recent years, much research focuses on

[1] http://www.alexa.com/siteinfo/weibo.com

G.A. Wang et al. (Eds.): PAISI 2013, LNCS 8039, pp. 19–28, 2013.

Micro-blog content classification [4-8]. Many researchers take machine learning approaches such as Naïve Bayesian (NB), Support Vector Machine (SVM) and K Nearest Neighbors (KNN) to mine content categories. The challenge is that classification effect may be hampered by the lack of semantics and content ambiguity. Sina-weibo provides us a natural way to gather posts of the same topic. It keeps the distinguishing characteristic of trending topics. Sina-weibo's trending topics are the special topics whose content are highly related to hot topics in our daily life. In the form of trending topics, we can group posts with similar semantics together conveniently. Although trending topics gather semantic-related posts to some extent, the ambiguity of Chinese language still prompts us to find more effective classification methods.

In this paper, we propose a novel framework to classify Micro-blog topics into different categories. We make full use of Sina-weibo's trending topics and Baidu Encyclopedia[2] to extract effective features. Then we employ these features and the Adpredictor-based model to classify test set topics into different categories. In our work, we generate the training set and test set using random sampling methods. We have conducted experiments on China's important Micro-blog platform Sina-weibo. The reliable classification results indicate the effectiveness of the proposed model.

The rest of this paper is organized as follows. We discuss related work in Section 2. In Section 3, we present our proposed framework and solution in detail. We discuss experimental setup and results in Section 4. Finally, we bring up the summarization and future work in Section 5.

2 Related Work

The key step in topic classification is how to make best use of short text corpus and generate features with high quality. There are many corresponding algorithms for this task, such as N-Gram, Bag-of-Words, vector space models and LDA-based topic models. These methods prove to be useful. However, due to the ambiguity of Chinese short text corpora, we need to bring up more suitable methods. The following threads of related work can provide us valuable research directions.

Researchers focus on classifying sentiment, topics and generating features with machine learning approaches. Lee et al. [5] proposed a Naive Bayes Multinomial classifier to classify the Twitter topics. The authors constructed word vectors with trending topic definition, tweets, and the commonly used tf-idf weights. Experiments on a database of randomly selected 768 trending topics showed that classification accuracy of up to 65% can be achieved using text-based classification modeling. Fan et al. [6] introduced a method of feature extension to classify Chinese short text based on Wikipedia. They first associated Wikipedia concepts based on statistical laws and categories, and then they adopted the KNN algorithm to train effective classifiers. Liu et al. [4] proposed a novel feature selection method based on part-of-speech and HowNet to extract these short text features. These methods are effective in providing useful word features.

[2] http://baike.baidu.com/

In this paper, to achieve reliable classification results, also inspired by the idea of taking advantage of external knowledge base [6, 8], we propose a novel classification framework using Baidu Encyclopedia and the Bayesian inference methods. First, we make full use of Sina-weibo's trending topics to gather semantic-related posts together. Second, we introduce Baidu Encyclopedia to expand corpus semantics and get keywords set. Third, we train efficient classifiers with the probabilistic graphic model-Adpredictor, which employs the Bayesian inference methods and classifies the test set with the highest probability label.

3 Proposed Methodology

In this section, we firstly define the problem of classifying Micro-blog topics with the predefined labels. Assume that we have gathered trending topics from Sina-weibo, with $\boldsymbol{T} = \{t_1, t_2,\ldots, t_n\}$ denoting trending topics list. Our task is to predict the labels of test topics set $\boldsymbol{T}^{\text{test}} = \{t_1^{\text{test}},\ldots, t_m^{\text{test}}\}$, given the labels of training set $\boldsymbol{T}^{\text{train}} = \{t_1^{\text{train}},\ldots, t_q^{\text{train}}\}$ and all the posts collection $\boldsymbol{P}= \{ p_1, p_2,\ldots, p_n \}$. In this task, we focus on Bag-of-Words features, aiming to train classifiers with high-quality words. We can transfer the classification problem into a category belonging prediction task. In the feature selection process, we extract the whole words set $\boldsymbol{D} = \{d_1, d_2,\ldots, d_k\}$ from the posts collection $\boldsymbol{P}$ and the training set $\boldsymbol{T}^{\text{train}}$. Then for each given test topic t^{test}, we generate its word feature vector $\boldsymbol{x} = \{ x_1, x_2,\ldots, x_k \}$ from $\boldsymbol{D}$, noting that $\boldsymbol{x}$ shares the same vector dimension k with $\boldsymbol{D}$, and each element x_i in $\boldsymbol{x}$ represents the corresponding weight of word d_i. We aim to learn the possibility t^{test} belongs to a predefined category with the denotation $p(y|\boldsymbol{x})$. Here, we denote the belonging/not-belonging state with $y \in \{-1,+1\}$, where 1 represents t^{test} belongs to the predefined category and -1 represents the not-belonging status.

3.1 Solution Framework

The framework of our proposed model is shown in Figure 1. We make a preliminary attempt to classify Micro-blog topics with basic word features, which integrates the trending topic posts and the external knowledge base. Our method first applies Baidu Encyclopedia to obtain the topic-related text and keywords set. Then it processes the Micro-blog posts to generate candidate features set. Next, we update micro-blog's features weight with the keywords and obtain the final classification features. In the end, we adopt the Adpredictor-based model to classify the test topic collection with the predefined category labels. We show the detailed steps as follows.

3.1.1 Introducing the External Knowledge Base

Posts of Sina-weibo own the characteristics of free written style and short lengthy. Therefore, it would be reliable to construct effective word features spaces with the semantic support of external knowledge bases. We choose Baidu Encyclopedia as our external knowledge base for its standard words use and plausible topic explanation. In this step, we first crawl the explanation pages for all the topics from the Baidu

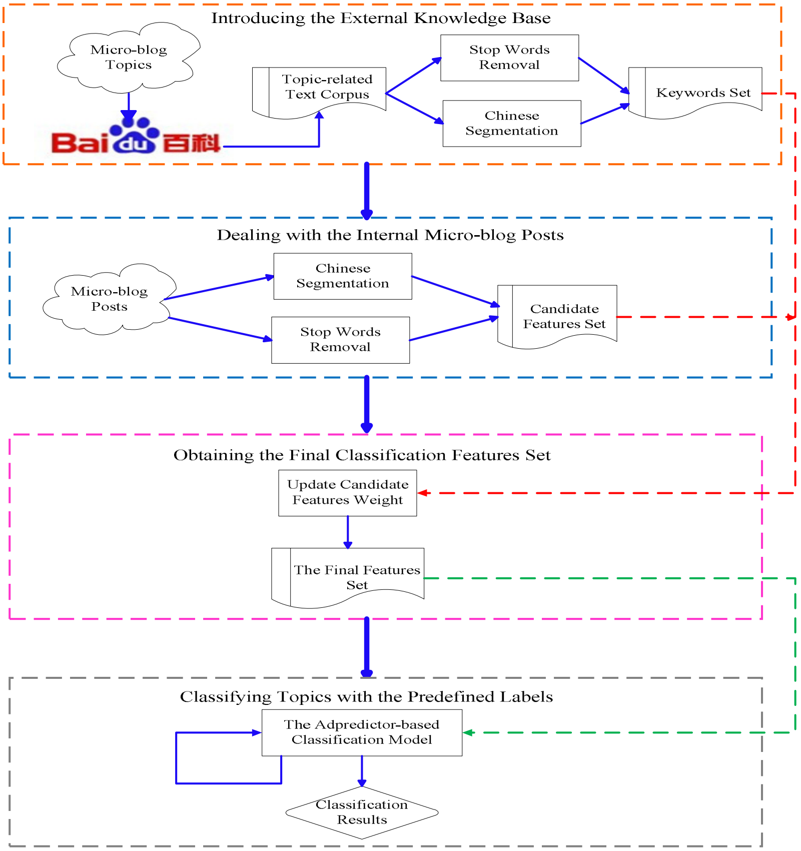
Introducing the External Knowledge Base
Micro-blog Topics
Baidu 百科
Topic-related Text Corpus
Stop Words Removal
Chinese Segmentation
Keywords Set
Dealing with the Internal Micro-blog Posts
Micro-blog Posts
Chinese Segmentation
Stop Words Removal
Candidate Features Set
Obtaining the Final Classification Features Set
Update Candidate Features Weight
The Final Features Set
Classifying Topics with the Predefined Labels
The Adpredictor-based Classification Model
Classification Results

Fig. 1. The Framework of Classifying Micro-blog Topics

Encyclopedia website, and then we extract the topic-related sentences on these pages as the external feature corpus. With the process of Chinese Segmentation and Stop Words Removal, we can get the segmented words. After calculating the tf-idf weight for each segmented word, we will get the ranked keywords list for all the categories.

3.1.2 Dealing with the Internal Micro-blog Posts

In Step 2, we gather posts of each topic together, then we generate candidate feature set with the module of Chinese Segmentation and Stop Words Removal. Notably, the generated candidate feature set is not ready for the classification model training. The existing of informal words and mistaken spelled ones may ruin the feature spaces, which will then hamper the classification performance. We need to use the technology of feature engineering to acquire the high-quality features set.

3.1.3 Obtaining the Final Classification Features Set

With the Baidu Encyclopedia keywords set, we can update the tf-idf weight of the candidate features set easily. Then we rank the updated word features set and obtain the final classification features that are most related to the predefined categories. Take the word “May Day” as an example, “May Day” is a popular singer group in China. When processing with the category “entertainment stars”, the Baidu Encyclopedia identifies “May Day” as a keyword, for its frequent co-occurrence with other popular entertainment-stars-related words. However, the tf-idf value of “May Day” may weight low in Micro-blog for its common use in people's daily conversation about dates. With the help of the Baidu Encyclopedia keywords set, “May Day” will stand its own way to distinguish the category “entertainment stars” together with other related features.

3.1.4 Classifying Topics with the Predefined Labels

In this step, we propose to employ the Adpredictor model to classify the test topics to the predefined category. This Adpredictor-based classification model shows its advantages in the following aspects. First, it reveals the relationship between the labels and their features in the form of probability dependency, which can then be graciously represented by the model parameters. Second, the linear training time and small memory demand ensures the model’s scalability to very large dataset.

3.2 The Adpredictor-Based Classification Model

In this section, we will describe our Adpredictor-based classification model in detail. Adpredictor has been well applied to many research fields, including CTR prediciton in advertising [9], browser switching detection in search engine [11]. Due to its good scalability, we successfully introduce Adpredictor into our classification framework. Adpredictor is a generalized linear model with a probit link function as follows:

$$p(y \mid \mathbf{x}, \mathbf{w}) := \Phi(\frac{y \cdot \mathbf{w}^T \mathbf{x}}{\beta}) \tag{1}$$

We introduce $\boldsymbol{w}$ to describe Micro-blog content features weight. By introducing the Bayesian inference methods, we build strong connections between the predefined categories and $\boldsymbol{w}$, and then features weight $\boldsymbol{w}$ could be estimated with the introduced factors. The probability inference mechanism proves to be valid in finding reasonable $\boldsymbol{w}$ values. The detailed derivation process can refer to [10]. After estimating content features’ weight vector $\boldsymbol{w}$, we need to infer the prediction equation. Given the prior probability distribution p(y|$\boldsymbol{x}$,$\boldsymbol{w}$) and $p(\boldsymbol{w})$, category belonging probability distributions can be derived as the integral:

$$p(y \mid \mathbf{x}) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p(y \mid \mathbf{x}, \mathbf{w}) \cdot p(\mathbf{w}) d\mathbf{w} \tag{2}$$

Therefore, for each given category, we utilize the training set features and the matching topic instances to train correspondent classifier. By adopting these classifiers, we predict the probability the test set topics belongs to different categories. Finally, we attach these topics with the most likely category label.

4 Experimental Results

In our paper, we perform classification experiments on Sina-weibo dataset to demonstrate the effectiveness of the proposed model. We evaluate the performance of our model with the detailed classification metrics on different categories. In order to prove the effectiveness of our approach, classification results without external knowledge base are introduced as the comparison of reference.

4.1 Dataset Description and Preprocessing

Our dataset is composed of Sina-weibo users' activity, which includes their content information about trending topics, with the time interval of 16 days from October 29th, 2011 to November 13th, 2011[3]. According to Sina-weibo's category list of trending topics[4], also considering researchers' valuable classification work [6,7], we predefine 10 categories, including variety shows, cartoon animation, movies&TV series, electronic games, entertainment stars, tourism, commercial brands, health, cars and sports. We set the training/test dataset ratio as 80%/20% with random sampling methods,

Table 1. The topic number & Baidu Encyclopedia keywords list on Sina-weibo

Categories	# Topic	Keywords Examples
variety shows	206	Mr. Burns, talk show, avenue of stars, Hunan Satellite TV, happy family
cartoon animation	172	Miyazaki, Astro, Sponge, hiroshi agasa, Death Note
movies & TV series	221	Lunmei Gui, Best foreign language film, Silent Hill, trailer
electronic games	163	Game of Life, adventure island , Online games, single player games, GPU
entertainment stars	239	Luodan Wang, offbeat boys, Jimmy Lin, Universal Artist Award, May days
tourism	158	Kekexili, beijing happy valley, Putuo, Malacca, Qinghai-Tibet
commercial brands	215	Vancl, Sea Fishing, CASIO, Outlet, Amazon
health	186	air quality, sub-health, WHO, flu vaccine, health times
cars	193	dirver, Dongfeng Citroen,double clutching, Toyota, car industry
sports	201	FC INTERNAZIONALE, AC Milan, Serie A, Australian Open, Open Tennis Championship

[3] http://open.weibo.com

[4] http://huati.weibo.com

repeat the sampling process ten times and take average metric scores as the final experiment results. We show the topic number and the generated Baidu Encyclopedia keywords in Table 1.

4.2 Metrics Evaluation

We empirically set the model parameter β as 5, we then verify the effectiveness of our model on the detailed metrics with Sina-weibo dataset. Table 2 shows us the comprehensive performance of our model on different metrics.

Table 2. The comprehensive performance on the detailed metrics

Categories	Precision	Recall	*F*1-score
variety shows	**0.913**	**0.825**	**0.867**
cartoon animation	0.792	0.823	0.807
movies & TV series	**0.868**	**0.844**	**0.856**
electronic games	0.752	0.791	0.771
entertainment stars	**0.956**	**0.871**	**0.912**
tourism	0.724	0.845	0.787
commercial brands	0.831	0.743	0.785
health	0.665	0.848	0.745
cars	0.783	0.897	0.836
sports	0.775	0.854	0.813

It can be seen from Table 2 that our model achieves satisfying classification results. Most of the categories achieves the *F*1-score higher than 0.75. As for the labels of “entertainment stars”、“variety shows” and “movies & TV series”, the classification performances are better than other labels, implying that trending topics of these labels possess rich semantic information and are easy to distinguish. Meanwhile, the classification results are consistent with the phenomenon that users of Sina-weibo tend to participate in the discussion of entertainment topics, which reflect Sina-weibo’s platform characteristic. On the other hand, although many topics can be classified reliably, some topics still need to improve their performance. The performance of health-related topics is unsatisfactory, the wide “health” category topic range may influence the generated features expressive ability. In our future work, we need to adopt more fine-grained feature engineering methods, so as to enhance the features quality. In order to verify the effectiveness of adding Baidu Encyclopedia, we make the comparison experiment. Table 3 shows us the detailed classification performance.

It can be seen from Table 3 that classification performances on these metrics drop when we remove the Baidu Encyclopedia keywords. The complementary classification results in Table 3 imply the significance of adding external knowledge base to enhance features’ quality and improve the classification accuracy. Besides, whether adding Baidu Encyclopedia external knowledge base or not, the category “entertainment stars” owns the highest classification accuracy. The reason may be that people on Sina-weibo talk much of entertainment stars, and their concern is more centralized than other kinds of topics.

Table 3. Classification performance without the Baidu Encyclopedia

Categories	Precision	Recall	$F1$-score
variety shows	0.813	0.757	0.784
cartoon animation	0.721	0.798	0.758
movies & TV series	0.752	0.746	0.749
electronic games	0.682	0.758	0.718
entertainment stars	**0.848**	**0.796**	**0.821**
tourism	0.675	0.770	0.719
commercial brands	0.761	0.749	0..755
health	0.606	0.758	0.674
cars	0.743	0.847	0.792
sports	0.717	0.805	0.758

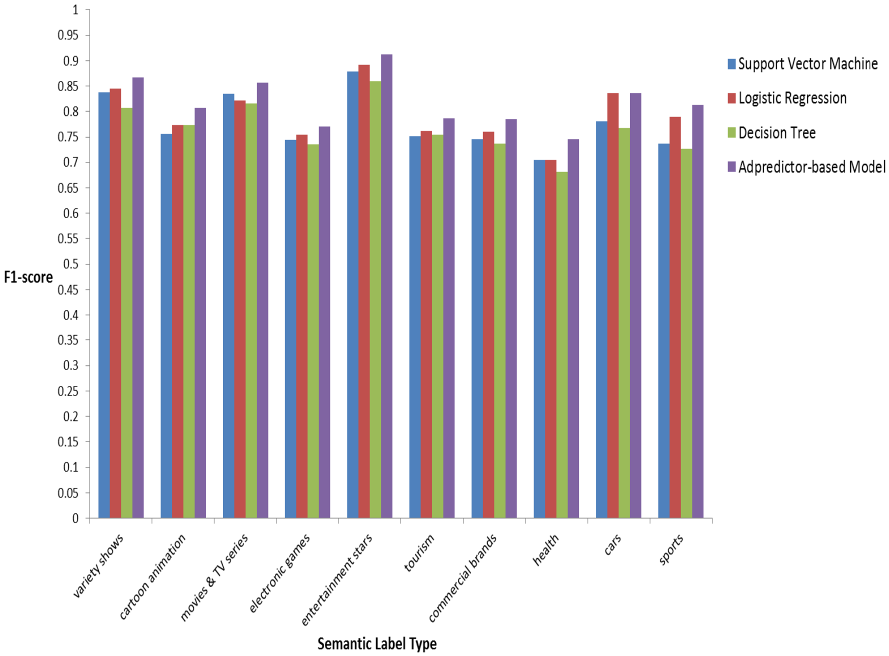

Fig. 2. The F1-score performance of various classification models

In order to verify the effectiveness of the proposed framework, we choose classical classification models Logistic Regression (LR), Support Vector Machine (SVM) and Decision Tree (DT) as the baseline models, and compare them with the Adpredictor based model. Figure 2 shows us the detailed F1-score performance of each classification model under various semantic labels.

In Figure 2, the F1-scores of the baseline classification models show similar trends under various semantic labels with the proposed model, they also achieve satisfactory performance on entertainment-related topics. These results demonstrate that the proposed model can mine users' main interest concerns on Sina-weibo reasonably.

Besides, when compared with other three baseline classification models, the proposed one improves the F1-score metric of all the semantic labels, which verifies the effectiveness of the proposed method when dealing with the micro-blogging short text classification problem.

5 Conclusions and Future Work

In this paper, we propose a novel framework to classify Micro-blog topics. In our framework, we first introduce Baidu Encyclopedia to extract effective features, and then we train efficient classifiers with the Adpredictor-based classification model. Experiments on Sina-weibo dataset demonstrate the effectiveness of the proposed approaches. The contributions of our work can be summarized as follows. First of all, the introduction of Baidu Encyclopedia offers new solutions to increase Chinese short text semantics. Secondly, the good classification results of Bayesian inference methods broaden our scope when handling with text classification tasks. Thirdly, with the meaningful labeled semantic topics, decision support departments will benefit from the semantically clear micro-blogging information and grasp the dynamics of public opinion better, thus maintaining social security and stability. In our future work, we will work on mining more effective micro-blogging content features, so that we could take more fine-grained classifying approaches to enhance our model performance and explore its application in social security area.

Acknowledgements. This research is supported by NSFC grants (No. 61172106, No. 71103080, No. 91124001), BJNSF (No. 4112062).

References

1. Zeng, D., Chen, H., Lusch, R., Li, S.-H.: Social media analytics and intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010)
2. Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.: The Role of Social Networks in Information Diffusion. In: Proceedings of the 21st ACM Conference on the World Wide Web 2012, pp. 519–528. ACM (2012)
3. Akshay, J., Xiaodan, S., Tim, F., Belle, T.: Why We Twitter: Understanding Microblogging Usage and Communities. In: Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM (2007)
4. Liu, Z., Yu, W., Chen, W., Wang, S., Wu, F.: Short text feature selection for micro-blog mining. In: International Conference on Computational Intelligence and Software Engineering (CiSE), pp. 1–4 (2010)
5. Banerjee, N., Chakraborty, D., Joshi, A., Mittal, S., Rai, A., Ravindran, B.: Towards Analyzing Micro-Blogs for Detection and Classification of Real-Time Intentions. In: Sixth International AAAI Conference on Weblogs and Social Media (May 2012)
6. Lee, K., Palsetia, D., Narayanan, R., Patwary, M., Agrawal, A., Choudhary, A.: Twitter trending topic classification. In: Proceedings of 11th International Conference on Data Mining, pp. 251–258. IEEE (December 2011)

7. Fan, Y., Liu, H.: Research on Chinese Short Text Classification Based on Wikipedia. New Technology of Library and Information Service 3, 47–52 (2012)
8. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using wikipedia. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 787–788. ACM (2007)
9. Graepel, T., Candela, Q., Borchert, T., Herbrich, R.: Web-Scale Bayesian Click Through rate Prediction for Sponsored Search Advertising in Micro-soft Bing Search Engine. In: International Conference on Machine Learning, pp. 13–20 (2010)
10. Herbrich, R., Minka, T., Graepel, T.: TrueSkill: A Bayesian Skill Rating System. In: Advances in Neural Information Processing Systems 20, pp. 569–576. The MIT Press (2007)
11. Gao, H., Li, Y., Li, Q., Zeng, D.: The Powerful Model Adpredictor for Search Engine Switching Detection Challenge. In: Workshop on Web Search Click Data, the Sixth ACM International Conference on Web Search and Data Mining (2013)