Classification of Chinese-To-English Translated Social Network Timelines using Naïve Bayes

Xiang-Ru Yu*, Zhong-Liang Xiang*, Dae-Ki Kang**

*Computer Software Institute, Weifang University of Science & Technology, Shouguang 262-700, Shandong, China

** Division of Computer and Information Engineering, Dongseo University, Busan 617-716, South Korea

yuxiangru1119@163.com, ugoood@163.com, dkkang@dongseo.ac.kr

Corresponding Author: Dae-Ki Kang

Abstract—**This study proposes a method that classifies Chinese social network positive-negative comments (Weibo) using naïve Bayes algorithm trained from English social network (Twitter) corpus. We train our text classifier using Twitter corpus (in English language), and use this classifier to classify Chinese text. In the previous research, Chinese sentences are processed using Chinese word segmentation algorithms before the application of machine learning algorithm. Chinese word segmentation algorithms split Chinese sentences into a series of words since a Chinese word consists of several Chinese characters unlike English sentences. Therefore, the quality of word segmentation algorithm obviously influences the accuracy of Chinese text categorization problems. In our research, we eliminate Chinese word segmentation stage (a traditional preprocessing stage of Chinese text classification) to avoid the effect on the quality of segmentation algorithms. Instead of Chinese word segmentation processing, we translate Chinese text into English text via Google translator. Based on Twitter corpus, we directly generate a text classifier by using naïve Bayes multinomial algorithm. Finally, the text classifier classifies a new Chinese text (a Weibo text, which has been translated into English by Google translation at preprocessing stage). We conduct an experiment comparing the performance of naïve Bayes multinomial algorithm and C4.5 in terms of accuracy.**

Keywords— **Text categorization, Classification, Naive Bayes, Multinomial model, Weibo, Comment**

I. INTRODUCTION

Weibo is a micro-blogging service in China. Weibo is a platform based on relationship among the users, aims to provide information sharing and collection. Weibo users can send 140 Chinese characters to update and share their information using Web, WAP and other client components. Weibo provides its service in a more free and swift manner for Internet users to communicate each other, declare their opinions, and record their emotion. In China, Sina Weibo[1] is a main micro-blogging service that has over 300 millions registered user accounts. The numbers of users sending messages in a day using Weibo is more than 100 millions.

Facing these huge amount of comments in Weibo, it is getting more difficult to categorize comments manually. Thus, we use machine learning method to deal with this text categorization problem.

We use naïve Bayes multinomial algorithm[1] to deal with Chinese Sina Weibo comments, and analyze the emotion of comments posted by Weibo users to decide if the user's comment is positive or negative. Due to the difference between Chinese language and English language, a special disposal to distinguish words is needed in the Chinese text preprocessing stage. The quality of Chinese word segmentation algorithms obviously influences the accuracy of Chinese text categorization problems. In this Chinese language text preprocessing phase, we propose a new method to use Google translator to translate Chinese text into English. Although the quality of translated result is not always sufficient for human understanding, but we believe the features for positive or negative emotions in the text should be reserved. We also believe that Google translator will produce translated text which generates statistical estimations in a reasonable quality for naïve Bayes multinomial algorithm which depends on the independence assumption of each word.

In this paper, first, we discuss the similarities and differences between the Chinese language and the English language in the text preprocessing stage. After that, we present naive Bayes multinomial algorithm. At last, we illustrate the experiment procedure and experiment results of the classification of Chinese Sina Weibo comments via naïve Bayes multinomial algorithm and C4.5.

[1] Sina Micro-blogging main website is `http://weibo.com`

The rest of the paper is organized as follows: Section II is about Chinese Language text preprocessing. In section III, we discuss naïve Bayes algorithm. In section IV, we present experiments and their evaluation. We draw a conclusion and future work at section V.

II. Chinese Language Text Preprocessing

Typically, we need a preprocessing phase for text classifier problem before training classifiers. Segmentation algorithms are needed for Chinese text preprocessing, stemming algorithms are needed for English text preprocessing, and morphological analysers are needed for Korean text preprocessing.

We summarize the similarities and the differences between the Chinese language and the English language in terms of the text preprocessing stage, given as follows:

The similarities between preprocessing Chinese text and English text:

- We need to remove all other elements except the words themselves to form a bag-of-words.
- We need to eliminate the uninteresting words, in other words, stop words [2].

Now we explain the differences between of them.

For Chinese text classification, we need Chinese word segmentation (CWS) algorithm, which is not needed for English text classification. It is because English text is naturally divided by space or punctuation into a single word. Therefore it is very easy to distinguish a single word in the text. However, in Chinese text, a sentence consists of Chinese characters. A Chinese character is the basic unit in Chinese text. It must be recognized and separated into one word from the narrative flow of Chinese characters. This Chinese word can be denoted as features and will be used at training stage of classifiers.

The qualities of word segmentation have great influence to Chinese text classification. There are some proposals including dictionary-based method, hidden Markov model (HMM) [3] and conditional random fields (CRF) [4] to deal with CWS.

On the other hand, English text preprocessing needs to deal with word morphology. For one English word, we need to reduce all probably emerging word morphology to one simple word prototype, but Chinese text does not need this kind of preprocessing.

In our paper, we translate Chinese Weibo text into English version via Google translator, and eliminate other elements except the words themselves from the text.

III. Naïve Bayes Algorithm

Naïve Bayes algorithm is one of supervised learning methods based on Bayes' rule on statistic theory scale, running on labelled training examples, and given by a strong assumption that all attributes in training examples are independent to each other given the class, so-called naïve Bayes assumption. It is generally believed that naïve Bayes assumption conflicts with the reality, but Domingos and Pazzani [5] have given some theoretical justifications that the binary independence assumption seldom harms effectiveness especially in a huge of training instances. On the other hand, naïve Bayes classifier shows high performance and rapid classification speed, which benefit from naïve Bayes assumption.

There are several kinds of naïve Bayes classification models. The most popular one is multi-variate Bernoulli model [6,7]. Another model is naïve Bayes multinomial model [8]. Detailed comparison and summary can be found in [1]. In our experiment, we use naïve Bayes multinomial model which is appropriate for text processing when a text is represented in terms of a bag of words.

We discuss Bernoulli model and the naïve Bayes multinomial model. In multi-variate Bernoulli model, a document is represented as a vector of binary attributes indicating which words occur and do not occur in the document.

Multinomial model concerns on post-probabilities of each word given by a certain class label. The post-probability of a word is one plus the number of occurrences of this word under a certain class in the scale of a certain document, and divided by the numbers of words in the vocabulary adding the sum of words under the class.

In naïve Bayes multinomial model, two assumptions are sometimes considered: the one is that the document length is independent of the class, another one is also independent between the probability of the word and the position of word appearing in the document.

Equation (1) is the definition of the post-probability of word given by the class under the naïve Bayes multinomial model:

$$P(w_t|c_j) = \frac{1+\sum_{i=1}^{|D|} N_{it} P(c_j|d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N_{is} P(c_j|d_i)} \qquad (1)$$

where c_j is the i-th class value that $c \in C = \{c_1, \dots, c_j, \dots, c_{|C|}\}$. The symbol V is a vocabulary that all the documents words map in, $|V|$ is number of the words in V. The symbol d_i denotes one document, and $d_i \in D = \{d_1, \dots, d_j, \dots, d_{|D|}\}$, $|D|$ is the number of the documents in D. The symbol N_{it} represent the occurred times of the word w_t in d_i, and $t \in \{1, \dots, |V|\}$.

We can see that:

$$\sum_{t=1}^{|V|} P(w_t|c_j) = 1 \qquad (2)$$

By given the class c_j.

On the other hand, if $\forall c_q \in C = \{c_1, \dots, c_{|C|}\}$ and none of the document d_i is marked with c_q, in this case, the post-probability of w_t given by the class c_q is shown as follows:

$$P(w_t|c_q) = \frac{1}{|V|} \qquad (3)$$

Equation (3) eliminates the influences of zero value post-probability.

IV. Experiments and Evaluations

In order to perform the experiments to classify Sina Weibo comments by naïve Bayes multinomial algorithm, we download data sets from the Twitter sentiment analysis projects [9]. Kouloumpis et al. [9] provide 8,000 comments which include 4,000 positive comments and 4,000 negative comments. This corpus gives us relatively sufficient confidence that which words are used in positive comments and which words are usually thrown in negative comments. To simplify matters, we restricted ourselves to binary classification problems.

The step of the experiment is as follows:

- Generate the classifier by running naïve Bayes multinomial algorithm on the twitter corpus.
- Test classifiers to the test data sets (Sina Weibo comments translated by Google translator)

We implement naïve Bayes multinomial algorithm on the twitter corpus and got the words post-probability. The word post-probability is shown as Figure 1:

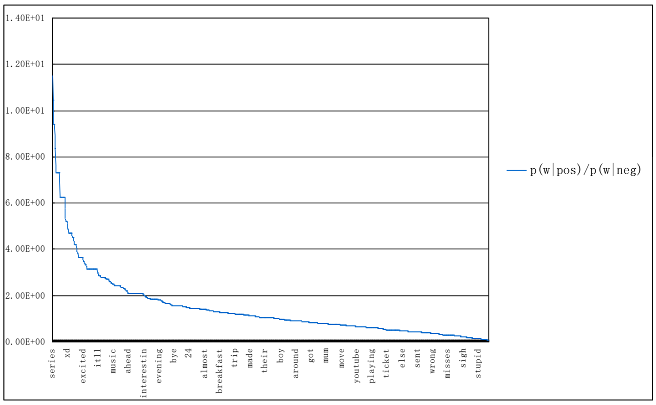

Figure 1. The formular *p(w|pos)/p(w|neg)* is the supporting ratio of words w_t appear in twitter corpus given positive class and negative class. The words which have high ratio show the more expression ability for the positive comments in Twitter, vice versa.

The test data set we have conducted experiments consists of 32 instances. We merge the training data set and the test data set into a single file, and choose 99.6 percentage split. The number 99.6 is based on the ratio of train instances and test instances in the merged file. We also perform C4.5 algorithm to this combined data set to compare the two algorithms in terms of accuracy, shown as Table 1.

TABLE 1. The accuracy information for compare the Naïve Bayes Multinomial algorithm and J48 by implement merged ARFF file.

Illustration	Multinomial	J48
Correctly Classified Instances	78.13%	78.13%
Incorrectly Classified Instances	21.88%	21.88%
Kappa statistic	0.5591	0.5591
Mean absolute error	0.2924	0.2974
Root mean squared error	0.3694	0.3863
Relative absolute error	58.48%	59.49%
Root relative squared error	73.8904 %	77.27%
Coverage of cases (0.95 level)	100%	100%
Mean rel. region size (0.95 level)	95.31%	96.88%
Total Number of Instances	32	32

V. Conclusion and Future Work

We have described experiments about Chinese text categorization by naïve Bayes multinomial algorithm. We have shown the details of naïve Bayes multinomial algorithm, and compared the experimental results between multinomial algorithm and C4.5. We have found the same ratio for correctly classified instances among the naïve Bayes multinomial algorithm and C4.5.

We also obtain some experience which except of enough test instances, there should be more to be considered for the text preprocessing. For instance, it will be useful to consider the case of elimination of meaningless words, the way to deal with numbers and symbols, and dealing with spelling mistakes.

In the future, we plan to design a classifier of combining the naïve Bayes multinomial algorithm, naïve Bayes algorithm and C4.5 algorithm by Adaboost algorithm to form a strong classifier, especially for large data sets.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A1A2013401).

Reference

[1] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," *AAAI-98 workshop on learning for text categorization*. vol. 752. 1998.

[2] Y. Wang, Various Approaches in Text Pre-processing, TM Work Paper No, 2004.

[3] W. J. Teahan, R. McNab, Y. Wen and I. H. Witten, "A compression-based algorithm for Chinese word segmentation," *Computational Linguistics*, vol. 26.3, pp. 375-393, 2000.

[4] T.-H. Yang, T.-J. Jiang, C.-H. Kuo, R. T.-H. Tsai and W.-L. Hsu, "Unsupervised overlapping feature selection for conditional random fields learning in Chinese word segmentation," in *Proc. of the 23rd Conference on Computational Linguistics and Speech Processing*, Association for Computational Linguistics, 2011.

[5] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, pp. 103-130, 1997.

[6] L. S. Larkey and W. B. Croft, "Combining classifiers in text categorization," in *Proc. of SIGIR*-96, 1996.

[7] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," in *Proc. of the 14th International Conference on Machine Learning*, 1997.

[8] D. Lewis and W. Gale. "A sequential algorithm for training text classifiers", in *Proc. of SIGIR*-94, 1994.

[9] E. Kouloumpis, T. Wilson, and J. Moore. "Twitter sentiment analysis: The good the bad and the omg!," in *Proc. of ICWSM*, 2011, pp.538-541.

Xiang-Ru Yu received a science master degree in computer science at Ocean University of China in 2010 and a Bachelor of Science (BS) degree in computer science at Mudanjiang Normal University in 2003. Currently, she is a Lecturer at Weifang University of Science and Technology. Her research interests include data mining and machine learning.

Zhong-Liang Xiang is a candidate Ph.D. student in computer science at Dongseo University in South Korea. He received a science master degree in computer science at Ocean University of China in 2010 and a Bachelor of Science (BS) degree in computer science at Mudanjiang Normal University in 2003. His research interests include data mining and machine learning.

Dae-Ki Kang is a professor at Dongseo University in South Korea. He was a senior member of engineering staff at the attached Institute of Electronics and Telecommunications Research Institute in South Korea. He earned a Ph.D. in computer science from Iowa State University in 2006. His research interests include intrusion detection, security informatics, ontology learning, and relational learning. Prior to joining Iowa State, he worked at a Bay-area startup company and at the Electronics and Telecommunication Research Institute in South Korea. He received a science master degree in computer science at Sogang University in 1994 and a bachelor of engineering (BE) degree in computer science and engineering at Hanyang University in 1992.