The multilayer sentiment analysis model based on Random forest

Wei Liu[1], Jie Zhang[2]

[1]School of Automation Beijing University of Posts and Telecommunications Beijing, China

[2]School of Automation Beijing University of Posts and Telecommunications Beijing, China

[1]twhlw@163.com, [2]mollychang@126.com

KEY WORD: text sentiment analysis, multi-features multi-base-classifiers meta ensemble learning sentiment analysis model, machine learning, situational awareness

Abstract With the rapid development of the Internet, artificial intelligence has gain widespread concern. Under the background, as one closely related discipline sentiment analysis's relevant research work have also been expanded. First, the paper analy existing text sentiment analysis method, compare the effect of a variety of emotional classification trained by traditional machine learning model. Second, it introduce ensemble learning methods, use random forest as meta learning method train base classifiers which trained through different feature sets. Though the experiments concluded that: by using a different set of features and different base classifiers, the ensemble model can obtain significant promotion, so the paper propose a new model " MFMB-ME，Multi-Features Multi-Base-Classifiers Meta Ensemble Learning Sentiment Analysis Model".

I. INTRODUCTION

With the development of the Internet, how to use the Internet to achieve social development becomes a direction of thinking. Because of the Internet's high speed and interconnection, a variety of social software and web site has been greatly developed. Using the Internet, people can show their feelings, ideas, views, etc. The large amount of unstructured text often contains the emotion and the viewpoint of events and objects. Through the analysis of the emotional text, we can dig out the people's emotion, evaluation of products, and the opinion to the popular events. Whether it is for the government or enterprises, how to get the correct analysis of the emotional information becomes very important. So how to dig out the emotional information from the vast amount of unstructured text becomes a direction to explore and research. Natural language processing is an important direction in the field of computer science and artificial intelligence. Its popular research directions include: syntax error correction, structural information extraction, semantic understanding, machine translation, emotion analysis, etc. The text sentiment analysis focuses on the analysis of the text about the speaker's emotion. Text sentiment analysis involves many disciplines, such as linguistics, data mining, machine learning, etc. As a wide range of knowledge and technology, people have made great efforts, also they have gained much achievement. In the text sentiment analysis, the main technology is divided into two categories: one is combining the emotional dictionary and rule, according to the text's positive emotional words and negative emotional words to carry out the emotional classification; the other is the use of machine learning method, by selecting feature word of the text, and labeling the training set and testing set with those feature word, final training the classifier by using machine learning methods. At the beginning of twenty-first Century, a new machine learning algorithm based on classification tree was proposed by Breiman and Cutler. Its main idea is to improve the prediction accuracy of the model by collecting a large number of classification trees. The model has been experienced many times, and the results have proved its effectiveness in many experiments. An important characteristic of the random forest is its fast processing, especially in dealing with large data. In this paper, we do experiment about the emotional analysis of the text based on the random forest as the training method. At the same time, the training model can calculate the importance of all the features, the paper studies the importance of different features of text sentiment. In natural language processing, word, stem, phrase all are the basic feature of the text. Most of the text classification systems use several basic features as training

feature of classifier to do the text processing tasks. In the near research, researchers have used the neural network to train language model, at the same time they obtained a distributed representation of the word in the fixed dimension. Bengio et al in 2001 have used a three layer of neural network to construct the n-gram language model, and achieved a better result than the ordinary n-gram[3]. On the basis of using the basic features of the text, this paper adds the word vector as the basic feature and do the experiment to explore the features' effect on the emotion analysis.

In this paper, it compare the difference of the result of text sentiment classification by traditional machine learning, a single feature set, multi-feature sets of meta-learning multiple classifiers ensemble learning. Experiments of traditional machine learning use decision trees, support vector machines, logistic regression and other methods, also compare results of classification performance by different traditions classification machine learning methods; use random forests as ensemble learning method train classifier based on a single feature set and analysis the classification performance; multi-features-classifiers meta ensemble learning method use the different combination of different text feature set (including word, stem, part of speech, grammar, ngram etc.) and different base classifier (logistic regression, language models, etc.) train classifier by random forest as meta-learning method the integrated, analysis the classification performance by different combination strategies.

The main innovations of this paper are: ①analysis existing text sentiment analysis method, compare the effect of a variety of emotional classification trained by traditional machine learning models;②introduce ensemble learning methods, use random forest as meta learning method train base classifiers which trained through different feature sets, propose a new model " MFMB-ME，Multi-Features Multi-Base-Classifiers Meta Ensemble Learning Sentiment Analysis Model".Though the experiments concluded that: by using a different set of features and different base classifiers, the ensemble model can obtain significant promotion. The structure of this paper is as follows: the second part is about the related work. The third part introduces the design of the model. The fourth part describes the design of the experiment and the analysis of the results.

II. RELATED WORK

A. Random Forest

Random forest is composed of many decision trees, and there is no association between each decision tree. In the process of generating random forest model, each decision tree is generated by random sampling, random sample set and random feature set. Each decision tree sum up the classification method by learning from a specific data, and the random sampling can ensure that there are duplicate samples can be classified by different decision tree, by this can be different decision tree classification ability to make evaluation.

Random forest model training process:

1) The training set as S, the testing set as T, features' dimension as F;

2) Randomly select sample from S as training sample S(i), the decision tree's training is start from the root;

3) If the termination condition is reached on the current node, set the current node as leaf node, the predicted output is the average of all samples' value on current node. Then continue training other nodes. If the current node does not reach the termination condition, randomly selected f-dimensional feature from the F-dimensional features without replacement. Use the f-dimensional features to look for one feature as k which can reach the best classification and set the corresponding threshold as 'th', the samples on the current node is divided into the left node if its value is less than the threshold, and the rest is divided into the right node.

4) Repeat step (2) (3) until all nodes have been trained or marked as leaf nodes.

5) Repeat step (2) (3) (4) until all regression trees have been trained.

6) Random forest regression model is made from regression trees, and the effect of the regression is evaluated by the residual mean square of the text data.

III. SENTIMENT ANALYSIS MODEL DESIGN

A. MFMB-ME

Model ‘MFMB-ME’ is divided into four levels, each layer corresponds to different modules, corresponding to different problems, they are: preprocessing module, features combination model, features preprocessing module, ensemble classification module.

(1) Preprocessing module: preprocess Raw text data, through Stanford's text processing tools acquired words, stem, part of speech, syntax and so on;

(2) Features combination model: combine different basic language features to obtain the complex language features, the different combination forms including the n-gram of same feature and the combination of different features;

(3) Features preprocessing module: use machine learning method to obtain meta-classifier. Meta-classifiers are mainly based on logistic regression, language model, ranking model;

(4) Ensemble classification model: use random forest to ensemble the meta-classifiers’ training result and train final classify model.

IV. THE DESIGN OF EXPERIMENT AND RESULTS ANALYSIS OF EXPERIMENT

A. Experimental data

The experimental data are emotional statements published in the social network, a total of 3000 emotional statements, they are divided into training set (64%), the validation set (16%), the test set (20%);

Table 1 Experimental data

Experimental data	Model training		Model testing
	Training data	Validation data	Test data
Sample	1920	480	600

B. Preprocessing module

Through the data preprocessing, we will get the basic features of the text, through word segmentation, stemming, grammatical processing using Stanford's text processing tools, we will get word, stem, grammar and other characteristics of the text.

C. The design and implementation of the experiment

(1) Experiment 1

Compare the effect of a variety of emotional classification trained by traditional machine learning models.

Table 2 Experiment 1 result based on traditional machine learning models

Machine learning model	Correct rate
Logistic Regression	0.82
Decision Tree	0.83
Support Vector Machine	0.84

(2) Experiment 2

The different combination of one feature is preprocessed by single machine learning method to get meta-classifier, and the meta-classifier’s output will be the input feature of the random forest to train out a model.

Experimental procedure:

A: Complex feature generation: the different combinations of one kind feature as complex features which will be preprocessed by step B; The different combinations showed as Table 3;

B: Meta-classifier: use machine learning method to train meta-classifier;

C: Random forest model’s training: ensemble the meta-classifiers by random forest；

D: Show the experiment’s result, the classification effect was evaluated by correct rate, the experiment’s result show as Table 4;

Table 3 Ensemble classify model based on single feature set

Machine learning	Char-N-gram	Words	Stem	Part of speech	Syntax
Logistic Regression	Tri-gram, 4-gram	Word, Bigram	Stem, Bigram	Tag, Bigram	Syntax
Rank Model	Tri-gram，4-gram	Word, Bigram	Stem, Bigram	Tag, Bigram	Syntax
KneserNey-Language model	——	Word, Bigram	Stem, Bigram	Tag, Bigram	——

Table 4 Ensemble classify model based on single feature set

Machine learning	Char-N-gram	Words	Stem	Part of speech	Syntax
Logistic Regression	0.823590	0.861730	0.842413	0.707356	0.836316
Rank Model	0.823615	0.858974	0.841711	0.708210	0.823801
KneserNey-Language model	——	0.680351	0.688292	0.663939	——

E：Experiment’s result analysis:Through the results, preprocessing method based on word lead the best result and the worst is base on part of speech. Also the effect of the logic regression is better than the language model under the same condition. Because all the correct rate is greater than 0.5, we can learn that all the character of text, word stemming, part of speech, grammar are meaningful for the text sentiment analysis. Also the word which have not been processed contain the most abundant emotions because of the lack of information in the process of the word segmentation, part of speech, grammatical transformation of the text, so their classification effect is poor. At the same time, the word vector obtained by Word2vec also achieved good experimental results, which means that the word vector in the text sentiment analysis is a great significance, so we can make this point to try more method.

(3) Experiment 3

The different combination of several features is preprocessed by single machine learning method to get meta-classifier, and the meta-classifier’s output will be the input feature of the random forest to train out a model. Experimental procedure:

A: Complex feature generation: the different combinations of several kinds feature sets as meta-classifier’s input, the meta-classifier is trained by single machine learning method. The different combinations showed as Table 5;

B: Same as experiment 2’s (B-D) steps;

C: Experimental data analysis: The experiment 3’s result as Table 6, from the result we can learn that the combination of a variety of feature sets will be better than single feature set.

Table 5 Ensemble classify model based on a variety of feature sets

Machine learning	Word and stem	Word and part of speech	Stem and part of speech	Stem and part of speech under syntax
Logistic Regression	Words, Stem，Word_Bigram, Stem_Bigram	Words, Tag，Word_Bigram, Tag_Bigram	Stem, Tag, Stem_Bigram, Tag_Bigram	Syntax_Stem_Tag

Table 6 Ensemble classify model based on a variety of feature sets

Machine learning	Word and stem	Word and part of speech	Stem and part of speech	Stem and part of speech under syntax
Logistic Regression	0.876246	0.867183	0.862185	0.772132

(4) Experiment 4

The different combination of several features is preprocessed by a variety of machine learning methods.

Experimental procedure:

A: Complex feature generation: the different combinations of several kinds features as complex feature which will be preprocessed by step B; The features' combination show as Table 7;

B: Feature preprocessing: use a variety of machine learning methods to obtain meta-classifier.

C: Random forest model's training;

Table 7 Ensemble classify model based on a variety of feature sets and different meta-classifier

Machine learning	Feature set	Correct rate
Logistic Regression, Rank Model, KneserNeyLM	Char_ngram, Word, Stem, Tag, Syntax, ngram	0.905586

E: Experimental data analysis: Through exp

stemming, part of speech, grammar, words and so on, classifier's effect will greatly improve, also compared with the simple classifier the random forest classifier is higher integrated with better classification results.

Reference

[1] HUANG Xuanjing,ZHANG Qi,WU Yuanbin.2011 . JOURNAL OF CHINESE INFORMATION PROCESSING, 25:185-192

[2] Bo Pang, Lil lian Le e. Shivakumar Vaithyanathan "Thumbs up Sentiment Clasification using Machine Learning Techniques[C] Procedings of the Conference on Empirical Methods in Natural Language Procesing"(EMNLP),2002

[3] Bengio, Y., & Ducharme, R. (2001). "A neural probabilistic language model." NIPS 13

[4] Ronan Collobert. Jason Weston. "A unified architecture for natural language processing: deep neural networks with multitask learning."160-167

[5]Bengio,Y., &Sen´ecal, J.-S.(2003)."Quick training of probabilistic neural nets by importance sampling." AISTATS'03

[6] Okanohara, D., & Tsujii, J. (2007). "A discriminative language model with pseudo-negative samples." Proceedings of the 45th Annual Meeting of the ACL, 73-80

[7] Morin, Frederic, and Yoshua Bengio. "Hierarchical probabilistic neural network language model." In AISTATS, vol. 5, pp. 246-252. 2005.

[8] Mnih, Andriy, and Geoffrey E. Hinton. "A scalable hierarchical distributed language model." In,vol.5,pp. 1081-1088. 2009.

[9] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." ICLR (2013).

[10]Fu, Ruiji, Jiang Guo, Bing Qin, WanxiangChe, Haifeng Wang, and Ting Liu. "Learning semantic hierarchies via word embeddings." ACL, 2014.

[11] Hinton, Geoffrey, and Ruslan Salakhutdinov. "Discovering binary codes for documents by learning deep generative models." 3, no. 1 (2011): 74-91.