Detecting Opinion Polarities Using Ensemble of Classification Algorithms

Li YueYang\textsuperscript{1, a} and Yi Zhi Wang\textsuperscript{1, b}

\textsuperscript{1}Beijing Jiaotong University, China

Email: \textsuperscript{a}973289456@qq.com; \textsuperscript{b}yzwang@bjtu.edu.cn

Abstract. Text classification enables developers to track consumer’s reaction to e-commerce products. Such information, often expressed in the form of raw emotions, can be used to measure consumers' emotions and their emotional preference for commodities, so as to help future consumers to make choices. However, capturing and interpreting human emotions expressed in product reviews is not a trivial task. Challenges stem from integrated approach and optimal feature combination methods of different classifiers. In this paper, we present a ensemble framework of text classification. It can found that the most adaptitve feature sets for classifiers. An effective method of CRF model to process medium and long text is proposed, this method significantly improves the CRF model's ability to handle text, length of which is more than ten. As base learning algorithms, three classifiers are integrated to improve the efficiency of classification, which are Support Vector Machine (SVM), Conditional Random Filed(CRF), and Naive Bayes Multinomial(NBM). The experiments prove the effectiveness of our proposed method.

1. Introduction

With the rise of the Internet, the amount of Internet users has exploded. Online shopping is an important component of Internet applications. Online shopping users accounted for 79.9% of the total netizens, and online shopping present escalating tendency. But there are two major problems with online shopping: (a) users can only evaluate products through product descriptions and product reviews. (b)There is a wide range of merchandise intermingled on e-commerce platforms. The supervision and management measures of e-commerce platforms on merchants are not perfect. How to choose the right products from tens of thousands of comments is called the main difficulty faced by customers.

In a book [1], professor Minsky from MIT university in the United States explicitly proposed the emotional recognition problem of intelligent machines for the first time. In order to further effectively measure emotions, Picard (1997) formally proposed the concept of "emotional computing" in the book “Affective Computing”[2]. A task has been established whose goal is to determine the polarity of a user's opinion based on emotional calculation with given user comments[3]. Emotion computing is about production of human emotion, emotion recognition, emotional expression and impact of emotional factors to measure the computational science. It uses computer technology to recognize the information carrier (e.g. physical characteristics, words from the text, sound, video, images, etc.) and the polarity tendency of human emotions (good, derogatory sense), and the intensity of the relationship between the measure, then to promote the emotion of computational science, also to promote the development of text analysis, etc. It has been found that emotions are a segment part of intelligence,
and not separate from it[4]. McLean [5] also pointed out that emotional computing will be the next important breakthrough and development direction in the field of artificial intelligence. At present, there are two approaches to the text emotion analysis: (a) sentiment analysis Lexicon-based approach[6], (b)sentiment classification based on machine learning[7]. The research method based on emotion dictionary [8][9][10]mainly relies on the construction of the emotion dictionary and the rules of emotion calculation. However The problem of unlisted words could not be resolved. Due to the training set with manual labeling, research based on machine learning can achieve a relatively high classification accuracy[11]. For the first time, machine learning method was introduced into the field of emotion classification research and a marvelous number of experiments were conducted[3]. A new training SVM algorithm is proposed[12]. [13] it is proposed that classifier integration can solve the problem of sentiment analysis. A novel feature combination method is proposed[14], which combines unigram features and filtered emotion dictionary features based on The Information Gain. Comparing six widely-used classifiers(e.g. Surport Vector Machine,Logistic Regression,Decision Tree etc.) It proves that NBM and SVM algorithms are better. Ensembling three classifiers(i.e. SVM, ME, NB) proves that explicit feature selection algorithm on unigram can improve performance, and different classifiers have their own best adapted feature sets[15]. Three different integration techniques (fixed rules, weighted combination and meta-classification) are compared, and it is concluded that ensemble method based on the weighted combination is the most effective[16]. An effective feature extraction method ,Z- score and The Information gain, is proposed[17]. That proved Document Frequency in feature extraction is less accurate than The Information Gain[18]. Part-of-Speech(Pos) tag feature is very effective for SVM classification algorithm[19]. We use Conditional Random Field (CRF) [20] , Support Vector Machine (SVM) [21] and Naive Bayes Multinomial (NBM)[22] as the learning algorithms.

The rest of the paper is structured as follows. Introduction of three classifiers (e.t. SVM, CRF, NBM), feature selection and feature combination design of each classifier in Section 2. In Section 3, we present our proposed method for ensemble construction. Experimental results with detailed analysis are presented in Section 4. Finally, we conclude in Section 5.

2. Algorithm Design
The main work of this paper is information classification of subjective texts, which are twitter and reviews for products. The information is divided into positive and negative. The algorithm theory based on supervised learning is Support Vector Machines(SVM), Conditional Random Filed(CRF) and Naive Bayes Multinomial(NBM). Three classifiers are trained separately through extracting respective feature sets. Detailed explanation of the above mentioned algorithms as follows.

2.1. SVM-based Sentiment Analysis
Support Vector Machine(SVM) is introduced by Vapnik in 1995[23], who established a mapping from document to term. In 1999, a new algorithm Namely Sequential Minimal Optimization for training Support Vector Machine (SVM) is discovered [24]. The Sequential Minimal Optimization breaks the large quadratic programming problem into a set of smallest possible quadratic programming problems. This method is indispensable for dealing with large datasets. [24] The strategy of polychotomous classification is proposed and application of SVM is mentioned. The advantages of SVM are suitable for small sample scenarios, nonlinear and high position pattern recognition. The toolkit used in this article is LIBSVM. Designed and developed by Professor Lin Zhi ren from Taiwan University.

2.1.1. Feature sets. We selected following five types of text features for the SVM classifier.

Feature 1: Part-of-Speech(PoS)
PoS information of the token pro- vides useful evidence to identify Emotional words. The potential affective words are adjectives, verbs, nouns, adverbs, interjections and prepositions. We use PoS tags of token as feature.
Feature 2: Sentiment Words
the intensity of emotional words and emotional score of sentence.
Feature 3: Number of emotional words
The number of positive words and negative words.
Feature 4: Privative words
Whether privative words appear before token.
Feature 5: Degree adverb words
Whether Degree adverb words appear before token.
Feature 6: The number of Special symbols
The number of question marks and exclamation marks.

2.1.2. Feature extraction. The feature of privative words and degree adverb words needed to be used in conjunction with the feature of emotion words. The same token, whether there are privative words or degree adverb words in front of them expresses different emotions. For example, this dress is not beautiful. The token beautiful is a positive word. However, there are privative words in front of the token, so the token of beautiful is a negative word in here. In the same way, the affective intensity of positive and negative emotion words is 1 and -1 respectively. For example, she is very cute. The intensity of cute is 1, yet the intensity of “very” is 2. And emotion intensity of cute then becomes 2. The feature of emotion score of sentence is the sum of the intensity of all the emotional words in the sentence.

2.2. CRF-based Sentiment Analysis
Conditional Random Filed was(CRF) is proposed by Lafferty et al in 2001[25]. Then it began extensive research on CRF [26] [27] [28]. The CRF based on lexical and syntactic has the best performance [29]. CRF considers the dependence between words. Compared with other classifiers, CRF has a strong ability of feature fusion and can combine multiple features. CRF is commonly used in tasks of sequence labels. Sentiment analysis based on text is to judge the emotional tendency of the whole sentence, and it’s not a typical label task. In order to turn sentiment analysis based on text into labels task, the polarity of the text corresponds to the polarity of words in the text. The CRF label task works best when the text length is less than 10[30], i.e., CRF is only more suitable for short texts. However, in reality, the text of the data set is often different. In order to make CRF suitable for long text. We proposed a new emotional judgment calculation method for CRF. If most of words in the text is positive, marking the text as positive. If most of words in the text is negative, marking the text as negative.

The feature of CRF is one more “word” feature than SVM.
Feature 7: “Word feature” is surface forms of tokens. The word itself is the direct basis and feature for discriminating emotional tendency. Because the words tend to be emotionally inclined, such as happy, sad, and annoyed.

2.3. NBM-based Sentiment Analysis
McCallum A [31] introduce Naive Bayes Multinomial(NBM). Six classifiers commonly used are researched and it proved NBM perform the best accuracy and use the least amount of time [14].

When the feature is discrete, polynomial model is used. we choose the feature is the word, which is the discrete feature. As in equation(1), \( N_{ya} \) represents the total number of words of class “a” in the training set. \( N \) is the total number of words in the training set. \( P(y_a) \) represents Prior probability. As in equation(2), \( N_{ya,x_i} \) is the sum of the occurrences of \( x_i \) in various documents under class “a”. \( |m| \) is the number of types of words in the training set. \( \alpha \) is smooth index. \( P(x_i|y_a) \) is posteriori probability. When \( \alpha \) is equal to 1, it named “Laplace smooth”. When \( \alpha \) is equal to 1, it is “Laplace smooth”. When \( \alpha \) is greater than 0 and less than 1, it is named “Lidstone smooth”. When \( \alpha \) is equal to 0, it no have smooth. But when a feature \( x_i \) does not appear in the training set, \( p(x_i|y_a) \) equals 0, leading to the posterior probability is 0,So we set \( \alpha \) to 1.
\[ P(y_a) = \frac{N_{y_a}}{N} \]  \hspace{1cm} (1)

\[ P(x_i \mid y_a) = \frac{N_{y_a \times i} + \alpha}{N_{y_a} + |m|} \]  \hspace{1cm} (2)

3. Proposed Method

In this section, our proposed framework is described. The entire process can be divided into the following steps.

Identify and implement the most relevant feature sets for SVM and CRF sentiment classifiers with The Information Gain.

Ensemble construction using the majority voting and the weighted voting to combine SVM, CRF and NBM. and find out the better performance to predict test set.

We use three classifiers namely SVM, CRF and NBM. The schematic diagram of the proposed method is depicted in Figure 1. The feature extraction is carried out by input training set. The features we use include lexical, syntactic as well as semantics information. We use information gain to find out the optimal combination of features for classifiers SVM and CRF respectively. We choose to optimize our evaluation indicators which are accuracy, precision, recall and F-messure. The evaluating indicator is maximized to find the optimal feature sets for classifier CRF and SVM. The optimal model SVM and CRF were obtained through the training of optimal feature sets. The above two models are ensemble with NBM based on the majority voting and the weighed voting respectively. Comparing the two ensemble methods and choose the better performance to predict test sets. The best ensemble is obtained by optimizing accuracy and F-messure.

![Figure 1](image-url) This figure depicts The schematic diagram of the proposed method.
4. Experiment

4.1. Datasets
The benchmark datasets of NLPCC task2 shared task is used for our experiments. Training dataset comprise of 5000 positive user reviews and 5000 negative reviews. User reviews in test dataset set counts to 2500. Brief statistics of the datasets are shown in Table 1.

| Dataset  | Positive | Negative | Total |
|----------|----------|----------|-------|
| training | 5000     | 5000     | 10000 |
| test     | 1250     | 1250     | 2500  |

4.2. Preprocessing
At first the datasets are preprocessed to remove the Hyperlink tags, email address and @. A authoritative Chinese word segmentation system namely NPLIR is used for word segmentation and part of speech tagging. After word segmentation, stop words were removed with the stop list of Harbin Institute of Technology.

4.3. Experimental Results
As mentioned above, we use three classifiers namely CFR, SVM and NBM. For implementation of CRF and SVM algorithms we use CRF++0.58 and LIBSVM. The evaluation index used in the experiment are accuracy, precision, recall and F-measure.

4.3.1. SVM Experiment Results Analysis
Results of SVM model are reported in Table 2. Comparisons between the results of different feature sets show that we can achieve better performance if we are able to find out the most appropriate set of feature. The feature combination of emotional words and privative with specific symbols had the best performance. The accuracy rate and F value reached 70.6% and 74.35% respectively. Among all the features, sentiment words play the most important role. When only sentiment words are used as features, the accuracy rate reaches 62.7%, and the F value is 71.87%. Privative and special symbols also play a certain role. When the features of private and special symbols are added, the accuracy rate increases by 6.9% and 1% respectively. Moreover, privative have a significant effect on the improvement of F value. However, when features of Pos, degree adverbs and number of sentiment word are added, The accuracy rate and F value are reduced to different degrees. It indicates that features namely PoS, degree adverbs and number of sentiment word are not suitable for SVM classifier model.

| Feature set | Accuracy | Precision | Recall  | F-measure |
|-------------|----------|-----------|---------|-----------|
| F1          | 51       | 52.14     | 52.34   | 51.66     |
| F2          | 62.7     | 59.61     | 84.18   | 71.87     |
| F1+F2       | 58.1     | 60.31     | 53.12   | 55.5      |
| F1+F2+F3    | 55.9     | 58.27     | 48.83   | 52.26     |
| F2+F4       | 69.6     | 65.52     | 85.74   | 76.83     |
| F2+F4+F6    | 70.6     | 68.6      | 78.52   | 74.35     |
| F2+F4+F5+F6| 58.5     | 56.22     | 85.55   | 69.49     |

4.3.2. CRF experiment results analysis
Results of CRF model are reported in Table 3. The results showed that word and sentiment polarity of words were the most suitable features for the CRF model. When using word and sentiment polarity of words features, the accuracy and F value reached 80.5% and 78.2% respectively. The performance was the best. However when features, such as Pos,
privative, specific symbols and degree adverb, are joining in, both accuracy and F value are reduced. So it indicates that the above four features are not suitable for the CRF model.

Table 3. Experimental results of different feature sets in CRF model.

| Feature set     | Accuracy | Precision | Recall  | F-messure |
|-----------------|----------|-----------|---------|-----------|
| F2+F7           | 80.5     | 84.38     | 75.97   | 78.17     |
| F2+F7+F1        | 51.8     | 51.9      | 79.88   | 62.84     |
| F2+F7+F4        | 80       | 83.47     | 75.97   | 77.93     |
| F2+F7+F6        | 52.5     | 51.93     | 97.07   | 68.14     |
| F2+F7+F5        | 51.5     | 51.38     | 97.85   | 67.48     |

4.3.3. Results of NBM model are reported in Table 4.

Table 4. Experimental results of NBM model

| Model | Accuracy | Precision | Recall | F-messure |
|-------|----------|-----------|--------|-----------|
| NBM   | 79.5     | 84.30     | 73.58  | 76.42     |

4.3.4. Results of classifiers Ensemble. Three classifiers namely SVM, CRF and NBM are chosen for constructing ensemble framework. In order to find the best ensemble construction, we use the majority voting and the weighted voting techniques. Evaluation results are reported in Table 5 that show the majority voting is the same as the weighted voting and accuracy and F values were improved 1.9% and 2.3% separately, respectively. It proves the effectiveness of our proposed method.

Table 5. Experimental results of ensemble based on the majority voting and the weighted voting

| Ensemble method | Accuracy | Precision | Recall | F-messure |
|-----------------|----------|-----------|--------|-----------|
| Majority voting | 0.824    | 0.857     | 0.787  | 0.805     |
| Weighted voting | 0.846    | 0.867     | 0.809  | 0.837     |

5. Conclusion
In this paper we have proposed a new ensemble learning for the sentiment analysis. As base machine learning algorithm, we choose SVM, CRF and NBM. Firstly, we find out the best feature sets for classifiers. The algorithm is based on the Information Gain. Maximizing some classification quality measures like accuracy and F-messure, so it will yield three models, each of which is the best model for the classifier. Secondly, the ensemble method combines three classification models mentioned above. Integrating classifiers based on the majority voting compare to the way based on the weighted voting. The ensemble method that performs better will be applied to predict test dataset. The weighted voting proves more effective by experiments. Experiments on benchmark datasets of NLPCC task2 prove our proposed method gets the optimal performance for the sentiment analysis. We compared experimental results with performance of feature sets. Under all the conditions, our proposed method proves its effectiveness. The main contributions of this work as bellow: (i). Building one new ensemble method for text classification that achieves the more excellent performance. (ii). Finding out the best combination of features suitable for the SVM and CRF classifiers. (iii). An effective method of CRF model to process medium and long text is proposed.

The current work focuses on finding out the outstanding way of features combination and ensemble methods for three classifiers. In future we would like to explore how to optimize more classifiers.

Reference
[1] Minsky M. The Society of Mind[J]. Personalist Forum, 1987, 3(1):19-32.
[2] Picard R W. Affective computing[M]. 1997.
[3] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retrieval , 2008.2 (1-2):1–135

[4] BUI, Duy Duc An; REDDEN, David T.; CIMINO, James J. Is Multiclass Automatic Text De-Identification Worth the Effort?. Methods of Information in Medicine, 2018, 57.04: 177-184.

[5] McLean A. Emotional intelligence is the future of artificial intelligence: Fjord. 2017. http://www.zdnet.com/article/emotionalintelligence-is-the-future-of-artificial-intelligence-fjord/

[6] YANG, Hui, et al. A hybrid model for automatic emotion recognition in suicide notes. Biomedical informatics insights, 2012, 5: BII. S8948.

[7] HASAN, Mehedi, et al. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. Journal of biomedical informatics, 2016, 62: 21-31.

[8] KESHI, Ikuo, et al. Semantically Readable Distributed Representation Learning and Its Expandability Using a Word Semantic Vector Dictionary. IEICE TRANSACTIONS on Information and Systems, 2018, 101.4: 1066-1078.

[9] WU, Liang; MORSTATTER, Fred; LIU, Huan. SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. Language Resources and Evaluation, 2018, 52.3: 839-852.

[10] JHA, Vandana, et al. A novel sentiment aware dictionary for multi-domain sentiment classification. Computers & Electrical Engineering, 2018, 69: 585-597.

[11] Liu B. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012, 5(1): 1-167.

[12] Wagner J, Arora P, Cortes S, et al. Deu: Aspect-based polarity classification for semeval task 4[C]//Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). 2014: 223-229.

[13] WANG, Junjie, et al. An online semi-supervised P300 speller based on extreme learning machine. Neurocomputing, 2017, 269: 148-151.

[14] YANG, Ang, et al. Enhanced twitter sentiment analysis by using feature selection and combination. In: Security and Privacy in Social Networks and Big Data (SocialSec), 2015 International Symposium on. IEEE, 2015. p. 52-57.

[15] PANG, Bo; LEE, Lillian; VAITHYANATHAN, Shivakumar. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002. p. 79-86.

[16] SINGH, Jaspreet; SINGH, Gurvinder; SINGH, Rajinder. Optimization of sentiment analysis using machine learning classifiers. Human-centric Computing and Information Sciences, 2017, 7.1: 32.

[17] CASTRO, Dayvid W., et al. Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties. Applied Soft Computing, 2017, 61: 1160-1172.

[18] ALRABABAH, Saif A. Ahmad; GAN, Keng Hoon; TAN, Tien-Ping. Mining opinionated product features using WordNet lexicographer files. Journal of Information Science, 2017, 43.6: 769-785.

[19] Zainuddin N, Selamat A, Ibrahim R. Improving twitter aspect-based sentiment analysis using hybrid approach[C]//Asian Conference on Intelligent Information and Database Systems. Springer, Berlin, Heidelberg, 2016: 151-160.

[20] VAIL, Douglas L.; VELOSO, Manuela M.; LAFFERTY, John D. Conditional random fields for activity recognition. In: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems. ACM, 2007. p. 235.

[21] BOUADIENEK, Nesrine; NEMMOUR, Hassiba; CHIBANI, Youcef. Fuzzy integrals for combining multiple SVM and histogram features for writer's gender prediction. IET Biometrics, 2017, 6.6: 429-437.
[22] SALEH, Ahmed I.; AL RAHMAWY, Mohammed F.; ABULWAFA, Arwa E. A semantic based Web page classification strategy using multi-layered domain ontology. *World Wide Web*, 2017, 20.5: 939-993.

[23] VAN DER GEER, J.; HANRAADS, J. A. J.; LUPTON, R. A. The art of writing a scientific article. *J. Sci. Commun.*, 2000, 163.2: 51-59.

[24] Hastie T, Tibshirani R. Classification by pairwise coupling[C]//*Advances in neural information processing systems*. 1998: 507-513.

[25] Wang, Yanna, et al. "CRF based text detection for natural scene images using convolutional neural network and context information." *Neurocomputing* 295 (2018): 46-58.

[26] LIU, Shifeng, et al. A CRF-Based Stacking Model with Meta-features for Named Entity Recognition. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Cham, 2018. p. 54-66.

[27] Chen, Tao, et al. "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN." *Expert Systems with Applications* 72 (2017): 221-230.

[28] WANG, Yan, et al. Biomedical event trigger detection based on bidirectional LSTM and CRF. In: *Bioinformatics and Biomedicine (BIBM)*, 2017 *IEEE International Conference on*. IEEE, 2017. p. 445-450.

[29] MO, Hsu Myat; NWET, Khin Thandar; SOE, Khin Mar. CRF-Based Named Entity Recognition for Myanmar Language. In: *International Conference on Genetic and Evolutionary Computing*. Springer, Cham, 2016. p. 204-211.

[30] Chen, Y. W., Wang, J. L., Cai, Y. Q., & Du, J. X. (2015). A method for Chinese text classification based on apparent semantics and latent aspects. *Journal of Ambient Intelligence and Humanized Computing*, 6(4), 473-480.

[31] McCallum A, Nigam K. A comparison of event models for naive bayes text classification[C]//*AAAI-98 workshop on learning for text categorization*. 1998, 752(1): 41-48.