# An Effective Hybrid Model for Opinion Mining and Sentiment Analysis

Kai Yang, Yi Cai*, Dongping Huang, Jingnan Li, Zikai Zhou, Xue Lei
School of Software
South China University of Technology
Guangzhou, China
*Email: ycai@scut.edu.cn

***Abstract*—Sentiment analysis and opinion mining is a task to analyze people's opinions or sentiments from textual data, which is very useful for the analysis of many NLP applications. The difficulty of this task is that there are a variety of sentiments inside documents, and these sentiments have variety expressions. Hence, it is hard to extract all sentiments using a dictionary that is commonly used. In this paper, we construct the domain sentiment dictionary using external textual data. Besides, many classification models can be used to classify documents according to their opinion. However, these single models have strengths and weaknesses. We propose a highly effective hybrid model combining different single models to overcome their weaknesses. The experimental results show that our hybrid model outperforms baseline single models.**

## I. Introduction

With the development of applications like network public opinion analysis, the demand on sentiment analysis and opinion mining is growing [1]. However, the structure of textual documents is variable according to authors' idiomatic usage, and there are extensive expressions to describe the same sentiment. Besides, a sentence may have different sentiment from different perspective. For example, a sentence like 'A is better than B' may have positive sentiment from the perspective of A, but the sentiment seems negative from the perspective of B. Meanwhile, articles from the internet usually contain buzzwords that is not collected by the common-used sentiment dictionary. This brings difficulty for us to find out the sentiments of documents.

Traditional opinion analysis methods rely on the sentiment dictionary to find out the sentiment words and judge the sentiments of documents. They cannot deal with the newly-born buzzwords. Besides, different domain have different idiomatic expressions, which may contain different sentiment words. Since these buzzwords appeared on the internet first and there are forums associating with different domains, we construct the domain sentiment dictionary using data from the corresponding forums. Hence, the newly-born buzzwords and domain sentiment words will be collected in our dictionary.

Traditional way of opinion mining finds out entities from a document first, and then the nearby sentiment words will be associated to the entities. However, this way has a problem that the sentiment words may be not related to the nearby entities. Instead, they are related to the aspects of these entities. For example, in the sentence 'Honda's small car has low fuel consumption and small size', words 'low' and 'small' are not related to the entity 'Honda', but related to aspects 'fuel consumption' and 'size' respectively. In this paper, we propose a three-layer model, where sentiment words only related to aspects, and the sentiments of entities are related to the sentiments of their aspects. In the example above, it is obvious that aspects 'fuel consumption' and 'size' have positive sentiment, thus sentiment of the entity 'Honda' can be evaluated as positive by combining the sentiments of aspects.

To classify documents into positive or negative sentiments, many traditional classification algorithms can be applied, for example, Support Vector Machine (SVM) or Gradient Boosting Decision Tree (GBDT). These single classification algorithms have their own strengths or weaknesses. SVM performs poorly when the data are sparse, while GBDT tend to over-fitting a specify dataset. Hence, these weaknesses restrict the performance of each single models. In this paper, we use stacking approach, which has been widely used in ensemble learning, to combine SVM and GBDT together. This hybrid model performs better than the single models for it overcomes the weaknesses of each single models.

In this paper, we have the following contribution: (a) A domain sentiment dictionary is constructed which can deal with problems created by buzzwords or domain sentiment words. (b) A highly effective hybrid model is proposed to reach higher accuracy in sentiment classification. (c) Several experiments are designed to verify the effectiveness of our proposed model, and the experimental result shows that the hybrid model outperforms the single models.

## II. Three Layers Model

Traditional sentiment analysis methods are based on sentiment dictionary. They find out entities from sentences, and then find out sentiment words nearby these entities. By applying sentiment dictionary, sentiments of entities will be obtained. However, the sentiment words may be not directly related to the nearby entities. Instead, they are related to the aspects of these entities. For example, in the sentence 'Honda's small car has low fuel consumption and small size', words 'low' and 'small' are not related to the entity 'Honda', but related to aspects 'fuel consumption' and 'size' respectively. In this paper, we propose a three-layer model, where sentiment

words only related to aspects, and the sentiments of entities are related to the sentiments of their aspects.

Given a document, we find out entities words first, and then find out aspects of the entities. Secondly, sentiment words will be found to describe the sentiments of aspects. Finally, we obtain the sentiments of entities by combining sentiments of their aspects together. Our proposed three layers model can solve the problem caused by ambiguous words. For example, in car field, the word '(high)' may have two different sentiments. 'high fuel consumption' express a negative opinion, while 'high chassis' is positive. In our proposed three layers model, the sentiment words will be attached to the corresponding aspects. Hence, 'high' will be considered a positive sentiment words when it belongs to aspect 'chassis', but it will become a negative sentiment words for aspect 'fuel consumption'.

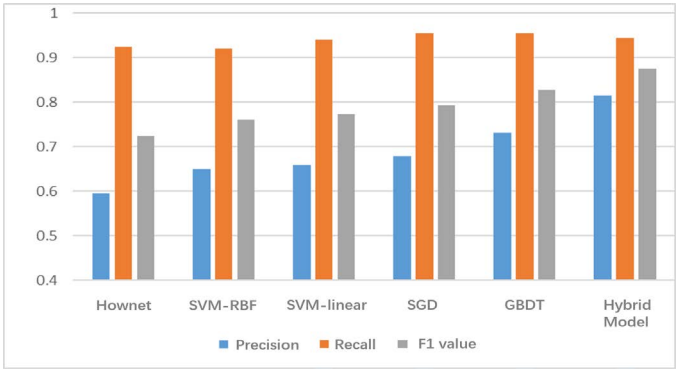


Fig. 1. Comparison of hybrid model with baseline models

## III. Hybrid Model

In this section, we propose a hybrid model combining SVM [2] and GBDT [3] together. The performance of GBDT is higher than that of SVM by 3-4%. After the comparison of GBDT and SVM, we can have the following conclusion: (a) SVM performs well when classifying sentences which have simple structure and strong opinion tendency, but it has poor performance for those complicated sentences. (b) GBDT performs well for long sentences with many sentiment words.

We use an example shown in Table I to make this clear. In this example, the first sentence has strong opinion tendency, thus SVM can get the correct sentiments. The third sentence is a long sentence, and it contains many sentiment words. In this case, GBDT gets the correct result, while SVM get the wrong result. Hence, these two model have both strengths and weaknesses. If these two model are combined together and overcome their own weaknesses, we can obtain a highly effective sentiment analysis model.

In this paper, we combine SVM and GBDT based on Stacking approach [4]. Stacking is an ensemble learning approach, which can be applied to combine different learning algorithms together to reach a better performance.

## IV. Experiments

In order to verify the effectiveness of our proposed hybrid model, we design the following baseline models for comparison.c (a) Baseline model 1: This model only applies a commonly used dictionary, called Hownet [5], to analyze sentiments; (b) Baseline model 2: This model only applies SVM model with radial basis function (RBF); (c) Baseline model 3: This model only applies SVM model with linear kernel function; (d) Baseline model 4: This model only applies SDG using Bagging as optimization method; (e) Baseline model 5: This model only applies GBDT.

TABLE I
Comparison of SVM and GBDT

| ID | Sentences | SVM | GBDT | Correct |
|---|---|---|---|---|
| 1 | 轩逸：天呐，我上榜了，我要发朋友圈去！ | pos | neg | pos |
| 2 | 【大事件】长安cs75－1.5t 劲越登场 | pos | pos | pos |
| 3 | 中控台上有着8 英寸的触摸屏也有着出色的质感，无论是触摸操作的反应速度还是显示的效果都对得上豪华二字，不仅如此，冠道为了兼顾各种使用情况特意将屏幕设计成多角度可调，无论多刺眼的阳光都不会影响使用 | neg | pos | pos |

The experimental results are shown in Figure 1. For single model, GBDT uses Boosting approach to integrate decision trees, but it tends to over-fit. From the experimental results, baseline model based on dictionary performs worst. The reason is that domain words in car field are not contained in Hownet. Besides, single SVM models also have poor performance, for the reason that the training dataset contains many sparse vectors. However, SVM with radial basis function (RBF) tends to over-fit, thus SVM with linear kernel function outperform that with RBF. The experimental result shows that the Hybrid model outperforms the baseline models.

## V. Acknowledgement

This work is supported by National Natural Science Foundation of China (project no. 61300137), Science and Technology Planning Project of Guangdong Province, China (No.2013B010406004), Tip-top Scientific and Technical Innovative Youth Talents of Guangdong special support program(No. 2015TQ01X633) and Science and Technology Planning Major Project of Guangdong Province (No. 2015A070711001).

## References

[1] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 342–351.

[2] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. ACM, 1998, pp. 159–168.

[3] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[4] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert systems with applications*, vol. 38, no. 1, pp. 223–230, 2011.

[5] Z. Dong and Q. Dong, *HowNet and the Computation of Meaning*. World Scientific, 2006.