# Combining Machine Learning and Crowdsourcing for Better Understanding Commodity Reviews

**Heting Wu[1], Hailong Sun[1], Yili Fang[1], Kefan Hu[1], Yongqing Xie[1], Yangqiu Song[2], Xudong Liu[1]**
[1]School of Computer Science and Engineering, Beihang University, Beijing,China, 100191
[2]Department of Computer Science, University of Illinois at Urbana-Champaign, United States
{wuheting, sunhl, fangyili, hukefan, liuxd}@act.buaa.edu.cn, xyqhello@gmail.com, yqsong@illinois.edu

## Abstract

In e-commerce systems, customer reviews are important information for understanding market feedbacks on certain commodities. However, accurate analyzing reviews is challenging due to the complexity of natural language processing and informal descriptions in reviews. Existing methods mainly focus on studying efficient algorithms that cannot guarantee the accuracy for review analysis. Crowdsourcing can improve the accuracy of review analysis while it is subject to extra costs and low response time. In this work, we combine machine learning and crowdsourcing together for better understanding customer reviews. First, we collectively use multiple machine learning algorithms to pre-process review classification. Second, we select the reviews on which all machine learning algorithms cannot agree and assign them to humans to process. Third, the results from machine learning and crowdsourcing are aggregated to be the final analysis results. Finally, we perform real experiments with practical review data to confirm the effectiveness of our method.

## Introduction

Nowadays, as more and more online purchases take place, ratings and reviews provided by consumers are widely used to analyze the market feedback of certain products. A rating is usually in the format of star level. For instance, a five-star represents the best and a one-star represents the worst. And a review is a certain length of text, possibly with pictures. The former is often regarded as the standard to classifying the latter. However, ratings can not always reflect consumers' feedback accurately. For example, a user can give a high-star rating even he is not very satisfied, in order to avoid the disturbance of corresponding customer service staff. Figure 1 shows the inconsistent customer evaluations using reviews and ratings respectively, in which the data is obtained from a major Chinese e-commerce company JD.com. The dataset contains 1,890 mobile phone reviews and ratings. Figure 1 (a) shows the percentage of positive, neutral and negative items with star based ratings. And Figure 1 (b) shows the manual analysis results with both ratings and reviews, which can be regarded as the ground truth. Therefore reviews can provide valuable information for understanding customers' evaluations.



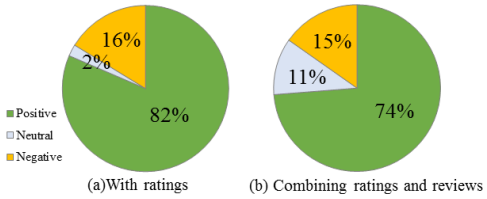


Figure 1: Inconsistent results with ratings and reviews with JD.com dataset

Straightforwardly, one goal of review analysis is to understand whether a review is positive, neutral or negative, which is essentially a problem of sentiment analysis. There have been a lot of studies(Hu and Liu 2004; Dalal and Zaveri 2014) using machine learning methods to analyze user reviews. Nonetheless, due to the complexity of natural language processing and the noise texts in reviews, the accuracy of existing methods still needs to be improved, especially for neutral reviews, which we will show in experiment section. In recent years, crowdsourcing (Liu et al. 2012) has been successfully used to deal with problems that are difficult for computer algorithms. However, crowdsourcing needs more costs for encouraging people to participate and low efficiency is also a challenging issue. This work aims at improving the accuracy of customer review analysis by combining machine learning algorithms with crowdsourcing. The core issue is to determine as less questions as possible to be processed with crowdsourcing, which is critical for the balance of accuracy, efficiency and costs. First, automatic reviews classification is done by several machine learning algorithms respectively. Second, reviews are selected to be assigned to humans to process, on which all machine learning algorithms do not agree. Third, we aggregate the results from machine learning and crowdsourcing to gain the final analysis results. With a dataset obtained from JD.com and a crowdsourcing platform (http://service4all.org.cn/crowdsourcing/ratings/login.jsp) developed by us, we perform an extensive set of experiments to evaluate the feasibility and effectiveness of our method.

## Method Description

As is shown in Figure 2, when users submit a request, *task manager* will crawl the corresponding product review, and

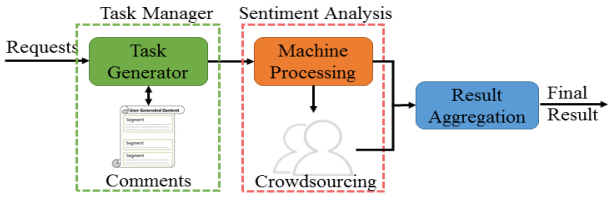


Figure 2: The workflow of review analysis

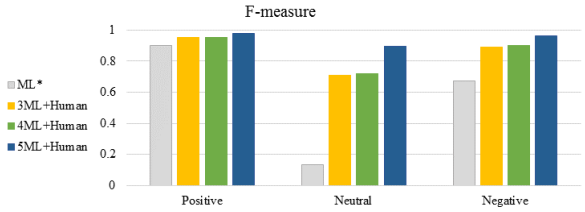


Figure 3: F-measure of classification results

do the transformation. Then, *sentiment analysis* module uses machine learning and crowdsourcing to process data and submit classification results to *result aggregation* module.

**Task Manager:** After users submit their review analysis requests, corresponding product reviews will be extracted for generating a series of tasks which will be delivered to the next module.

**Sentiment Analysis:** In this stage we combine machine learning with crowdsourcing. Determining whether a task should be done by computers or by crowd is the core issue, which directly influences the result accuracy and cost. To formally describe this problem, we denote the set of tasks as $T = \{t_1, \ldots, t_n\}$, and classify the tasks to 4 classes: positive(denoted as $G$), neutral(denoted as $N$), negative(denoted as $B$) and irrelevant(denoted as $I$). For a machine learning algorithm $f_i$ in $F = \{f_1, \ldots, f_n\}$, $f_i(T) = \{G_i, N_i, B_i, I_i\}$ represents its result set. Our problem, depicted as follow, is to determine which tasks, according to machine learning results $F(T) = \{G, N, B, I\}$, to be assigned to humans.

$$F(T) - (\bigcap_{i=1}^{n} G_i \bigcup \bigcap_{i=1}^{n} N_i \bigcup \bigcap_{i=1}^{n} B_i \bigcup \bigcap_{i=1}^{n} I_i)$$

This expression means that if the results of one task returned by all the algorithms are not all the same, the task will be delivered to the crowd.

**Result Aggregation:** Here we directly use majority voting to determine the crowdsourcing results and other aggregation methods can also be used. Then the crowdsourcing results will be combined with machine learning results.

## Experiments

**Experiment setups:** We crawl 9,738 reviews of 9 mobile phones from JD.com. The dataset is split into one training set (80%) and one testing set (20%). And we regard the crowdsourcing results as the ground truth. The experiments are performed as the following two groups:

- ML: We apply five machine learning algorithms including *Support Vector Machine(SVM)*, *Naive Bayes (NB)*, *K-Nearest Neighbor(KNN)*, *AdaBoost (AB)* and *Decision Tree (DT)* separately to classify the reviews.
- xML+Human: The $x$ denotes the top $x$ algorithms selected according to the accuracy from first group experiments. When the results of a task calculated by these $x$ algorithms are not the same, human's result will be accepted. Crowdsourcing experiments are performed by our platform mentioned above. We use F-measure to evaluate their performance.

**Results:** For the first group of experiments, we compute the average classification accuracies of all the five algorithms. The descendent ranking result is: *SVM*, *DT*, *KNN*, *AB* and *NB*. *SVM* achieves the highest accuracy of 0.817. The results of the second group experiments are shown in Figure 3. We also plot the result of *SVM* in Figure 3, which is represented as ML*. We can see that all the three hybrid methods performs better than the best of any single machine learning algorithm. In our methods, the cost of labor varies with the number of ML algorithms. And the accuracy rises as the labor costs increase. The most accurate one is the 5ML+Human method, with a F-measure of 0.979, 0.894 and 0.964 respectively for *positve*, *neutral* and *negative* categories. The 5ML+Human method has the most labor cost of 870 crowdsourcing tasks.

## Conclusion & Future Work

In this paper, we propose a hybrid method of machine learning and crowdsourcing for the sentiment analysis of commodity reviews. The core issue is to reduce the number of human tasks as best as possible, but to maintain the analyzing accuracy as well. Real experiments show our method is better than any single machine learning algorithm. In future, we will study the optimal combination of machine learning algorithms to further reduce the number of human tasks.

## Acknowledgements

This work was supported partly by China 973 program (No. 2014CB340304, 2015CB358700), partly by A Foundation for the Author of National Excellent Doctoral Dissertation of PR China (No. 201159) and partly by the Fundamental Research Funds for the Central Universities.

## References

Dalal, M. K., and Zaveri, M. A. 2014. Opinion mining from online user reviews using fuzzy linguistic hedges. *Applied Computational Intelligence and Soft Computing* 2014.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM.

Liu, X.; Lu, M.; Ooi, B. C.; Shen, Y.; Wu, S.; and Zhang, M. 2012. Cdas: a crowdsourcing data analytics system. *Proceedings of the VLDB Endowment* 5(10):1040–1051.