IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 21, 2019, accepted July 20, 2019, date of publication July 24, 2019, date of current version August 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2930905

Extracting the Pairs of Opinion Target and Opinion Term From Reviews With Adaptive Crowd Labeling

YUMING LIN[1], WEI ZHAO[1,3], YOU LI[2], HUIBING ZHANG[1], AND YA ZHOU[1]

[1]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
[2]Guangxi Key Laboratory of Automatic Detecting Technology and Instruments, Guilin University of Electronic Technology, Guilin 541004, China
[3]School of Data Science and Engineering, East China Normal University, Shanghai 200062, China

Corresponding author: You Li (liyou@guet.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61562014 and Grant U1711263, in part by the Project of Guangxi Natural Science Foundation under Grant 2018GXNSFDA281049, in part by the Research Project of Guangxi Key Laboratory of Trusted Software under Grant kx201916, and in part by the Project of Guangxi Key Laboratory of Automatic Detecting Technology and Instruments under Grant YQ17111.

ABSTRACT Labeling samples manually is a laborious, error-prone, and cost-consuming process, which is a traditional approach to preparing training data for supervised learning. Crowdsourcing provides an effective way to acquire labeled training data. In this paper, we propose an adaptive crowd labeling method to construct a set of the pairs of opinion target and opinion term iteratively under certain budget constraint. First, we assess the workers' reliabilities with a small number of labeled samples based on an EM process, since the worker's expertise varies widely on an open crowdsourcing platform. And then, the tasks are assigned to the high reliable workers for labeling the pairs of opinion target and opinion term. Finally, the responds of a sample from multiple workers are integrated to generate a final result based on the labelers' reliabilities as well as the dependency relation between opinion target and opinion term in this sample. A series of the experimental results show that the proposed method can achieve better extracting effectiveness compared with the baselines and the state-of-the-art methods.

INDEX TERMS Crowdsourcing, opinion target, opinion term, worker assessment, budget constraint.

I. INTRODUCTION

Online reviews have great reference value for potential customers, product manufacturers and service providers, since they include rich information on user experience and opinion. However, it is impracticable to investigate and analyze the users' opinion from massive reviews manually. Then, it brings urgent need for dealing with such tasks automatically and intelligently.

Opinion mining, also called sentiment analysis, targets at understanding users' opinion expressed in various medias like texts. In recent years, the fine-grained opinion mining has been given more and more attentions than the course-grained one like document-level, because it can provide more insight on user's opinion. Consider the review sentence shown in Figure 1, which is coming from a cell phone review. The solid boxes indicate the opinion targets, the dotted boxes indicate the opinion terms. The arrows mean the dependency relationship between the opinion target and the opinion term. It is clear that opinion targets and opinion terms include crucial information on user opinion expressed in review sentences. Therefore, extracting the pairs of opinion target and the corresponding opinion term in reviews is a core task within fine-grained opinion mining. We call this process as opinion pair extraction in this work, where the opinion pair is represented as a 2-tuple of ⟨*opinion target, opinion term*⟩. There are two opinion pairs ⟨*speed, amazing*⟩ and ⟨*sound quality, poor*⟩ in above example.

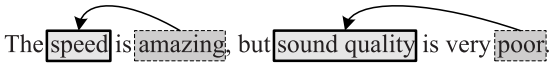


FIGURE 1. An example of opinion targets and opinion terms.

The supervised methods on extracting opinion pairs from reviews show relatively good effectiveness in existing works such as [1]–[3]. These methods need a large number of high quality labeled samples to train the extractor models. However, labeling samples manually is laborious, error-prone

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

and time-consuming. Crowdsourcing provides an effective way to construct training set for different learn algorithms by utilizing online collective intelligence. The prior *one shot*-based work on this problem collects multiple annotation results for all micro tasks released at once. The final result of a task is generated by some sophisticated models like Expectation Maximization-based method [4], Bayesian technique [5] and so on, since non-expert workers on crowdsourcing platforms have different annotation reliabilities due to their diverse education backgrounds and expertise. This type of methods tends to assess the workers' quality afterwards, which would lead to extra cost because all workers can be assigned the labeling tasks with equal chance.

Recently, some crowdscourcing platforms provide the task assignment mechanism of choosing specific workers for users, such as the *crowdspring*,[1] the *microWorkers*[2] and the *Figure Eight*.[3] Then, we can only assign tasks to the workers with high reliabilities rather than all workers in such scenarios. Inspired by this motivation, we propose an adaptive method to harvest the opinion pairs from reviews by crowd labeling, in which the opinion pairs are collected iteratively based on a forward assessment process on worker's reliability. Specifically, we assess the workers' reliabilities with a small set of labeled samples firstly. Only the reliable workers will be assigned the labeling tasks. And then we integrate the workers' responds to generate the final opinion pair(s) for each review sentence based on labelers' reliabilities and the dependence information of extracted opinion target and opinion term. Further, some results generated in previous iteration will be used to reassess the workers' reliabilities in next iteration, which can ensure the worker's reliability without extra cost.

The main contributions of our work can be summarized as follows:

1) We propose an adaptive crowd task assignment mechanism, in which the tasks are assigned iteratively based on the continuously updated assessment of workers' reliabilities with an EM process.
2) Since an extraction task will be assigned to multiple workers for ensuring the result's quality. We design a method to integrate multiple responds from workers based on the worker's reliability and the dependency information of opinion target and opinion term.
3) We develop a crowdsourcing system for harvesting the opinion pairs from review sentences. Experimental results indicate that our approach can improve the extraction performance significantly compared with several baselines.

II. RELATED WORK

Crowdsourcing is an effective way to prepare training samples for machine learning models [6]–[8], which involves two main steps generally: the design and distribution of task, the data integration of workers' responds. The former is responsible for decomposing the tasks into the wieldy micro-tasks and assigning these micro-tasks to workers with some certain mechanisms. Because the workers on crowdsourcing platforms prefer to perform simple tasks with just a little bit of effort and the overall cost is easier to control for users [9]. The latter focuses on integrating responds generated by workers into a final result for each task, which is a common way for assuring the result's quality.

A. THE DESIGN AND DISTRIBUTION OF TASK

The design and distribution of task is the foundation for solving the user's problem by crowdsourcing successfully. Jiang and Matsubata [10] decomposed a task into simpler sub-tasks based on the quality of the final results. Brambilla *et al.* [11] proposed a comparative, explorative approach for designing crowdsourcing tasks. This method defines a representative set of execution strategies, then executes them on a small dataset, and finally decides the strategy to be used with the complete dataset based on the quality measures for each candidate strategy.

Zhang *et al.* [12] used the entropy to model the task's informativeness, and proposed a probabilistic framework to select the most appropriate workers for a task. Guo et al. treated the task distribution as a recommendation problem [13], where both the workers' expertise levels and their interested points were integrated to recommend tasks based on each task's topic. Tunio *et al.* [14] further extended the interested points to payment, time and task type, and distributed a task to the workers with the highest matching degree by analyzing the effect weights of each factor.

The budget is often an important factor for labeling. Since many non-experts could attain an expert's annotation effect [15], the number of workers should be determined firstly by the task accuracy requirements [16]. Considering that the professional level of workers may have an improvement during the labeling process, a crowdsourcing task can be assigned several times to workers, and the workers can see others' annotations during each task release, but they may be interrupted by other workers' noisy answers.

As mentioned above, task publishers need to pay workers a certain amount of money to stimulate they complete the task. Li *et al.* [17] proposed an incentive mechanism for obtaining indoor location data, which probably takes the workers' privacy information into account, so it can be divided into two ways. One is fixed payment, which does not need to know the workers' privacy information. The other one can dynamically change the payment to workers by the value of feedback data and price fluctuation of different crowdsourcing platforms on the premise of knowing workers' privacy information.

B. DATA INTEGRATION OF WORKERS' FEEDBACKS

The annotations of crowdsourcing tasks are obtained based on workers' subjective judgment, so it often exists that some

[1] www.crowdspring.com
[2] www.microworkers.com
[3] www.figure-eight.com

workers' feedbacks are unreliable. Mitra *et al.* [8] addressed the challenge of obtaining high-quality annotations for subjective judgment oriented tasks of varying difficulty, and the experiments showed that the person-oriented strategy is superior to the process-oriented one. Because the former pays more attention to the workers' working status, it can improve the quality of workers' feedbacks by effective training.

There are also some problems with the person-oriented strategy. For example, due to the uncertainty of worker's reliability, it leads to a lot of noise in the workers' feedbacks. Such noise was verified experimentally to be harmful to the training data and trained model quality by Li and her colleagues [18], they further proposed a noise correction method called between-class margin-based noise correction (BMNC) for crowdsourcing [19]. Raykar *et al.* [4] designed a supervised learning algorithm based on Bayesian framework, which could extract valuable data from multiple workers' respond with noise. Donmez *et al.* [20] screened workers with high reliability and released tasks based on active learning, the labeling results of high-quality would be obtained finally by the minimum labeling effort. However, these two methods only focus on the impact of workers' professional level to the quality of feedback data, and lack the analysis of data itself property compared with our method.

Many traditional statistic methods can be used to integrate the workers' responds for generating the final results such as Bayesian model [5], probabilistic matrix factorization [21], Entropy [22], Markov model [23], Expectation Maximization [4] and differential evolution [24]. Whitehill *et al.* [25] put forward a classic data integration algorithm of workers' feedback, which combined the workers' reliability with the data itself property and iterates to produce high-quality feedback with the EM algorithm. Compared with our integrated method, Jacob's work did not take into account the budget and the change of the workers' reliabilities.

III. PROBLEM FORMALIZATION

Given a review sentence set $S = \{s_1, \cdots, s_m\}$ and a worker set $W = \{w_1, \cdots, w_n\}$. Let the symbol OP_i to be the set of true opinion pairs $op_i^j = \langle ot_i^j, ow_i^j \rangle$ included in review sentence s_i, where ot_i^j and ow_i^j are the corresponding opinion target and opinion term respectively. Notably, there might be multiple opinion pairs in a sentence.

We wish to extract the opinion pairs from these review sentences under a constraint of budget B on a crowdsourcing platform with n workers. In order to ensure the labeling result's quality, a labeling task would be assigned to multiple workers simultaneously. For simplicity and without loss of generality, we assume that the labeling cost is equal to 1 for all review sentences in this work. To reduce the complexity of opinion pair extraction task for workers, we provide them some opinion pair candidates extracted from the corresponding review sentence by some predetermined rules. For example, we assume that the opinion target should be noun or noun phrase, the opinion term should be adjective, verb or adverb. By this way, worker can select one or more opinion pairs from the candidates as the respond for a task. Then, the opinion pair extraction can be transformed into a classification problem for each review sentence.

Since a review sentence can be assigned to multiple workers for labeling, the responses of workers can be marked as the following matrix:

$$\hat{R} = \begin{bmatrix} \hat{r}_{1,1} & \hat{r}_{1,2} & \cdots & \hat{r}_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{m,1} & \hat{r}_{m,2} & \cdots & \hat{r}_{m,n} \end{bmatrix}$$

where the i-th row of $\hat{R}$ represents the workers' responses for review sentence s_i, $\hat{r}_{i,j}$ is a value from the predefined label set of s_i. At the same time, the matrix $\hat{R}$ also describes the task assignment, where $\hat{r}_{i,j} \neq 0$ means sentence s_i is assigned to worker w_j. It is worth noting that the matrix $\hat{R}$ is generated iteratively in our work, which is different from the methods of the *one shot* type.

In order to harvest the correct opinion pairs from the review sentence set as many as possible with a limited budget, we need to find an optimal labeling task assignment on m review sentences and n workers. Then, it turns to solve the following optimization problem:

$$R^* = \arg\max_R \sum_i \sum_j I(f(\bullet) = OP_i)$$
$$s.t. \sum_k \sum_t I(r_{k,t} \neq 0) \leq B \quad (1)$$

where I is the indicator function, $f(\bullet)$ is a fusion function for generating a derived result for sentence s_i based on the responds of multiple workers. For example, assuming sentence s_1 is assigned to worker w_1, w_2 and w_3 for labeling simultaneously. Then, the function $f(\hat{r}_{1,1}, \hat{r}_{1,2}, \hat{r}_{1,3})$ will generate a derived result $\hat{op}_1$ for sentence s_1.

IV. THE OVERVIEW OF THE PROPOSED METHOD

In this work, we propose an iterative adaptive crowd labeling method to harvest the high-quality opinion pairs form review sentences by crowdsourcing service. The overview of this method is shown in Figure 2, in which the solid lines with arrows describe the process flow and the serial numbers mean the process order.

Firstly, we assign all workers a small amount of labeled samples for labeling. Then, the reliabilities of workers are assessed based on the workers' responds on these samples. This assessment process will be discussed in Section V. By this way, the workers with low reliabilities are filtered, and the reliable workers will be assigned unlabeled samples for labeling. Notably, we only distribute a certain percentage of unlabeled samples to these reliable workers rather than all of them, because we will keep assessing workers' reliabilities throughout the labeling process. Next, the responds of workers are integrated to generate the final result for each sample, since each sample will be assigned multiple workers

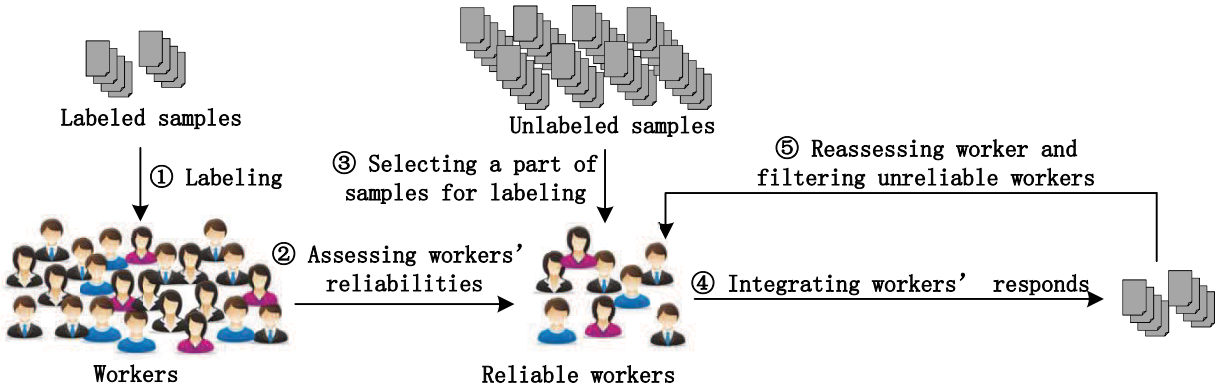


FIGURE 2. The overview of the proposed method.

for labeling. The generated results are used to reassess the workers subsequently and the unreliable workers will be filtered. So far, an iteration process of extracting opinion pairs is finished. At last, a certain percentage of unlabeled samples are assigned to the left workers for labeling, and these processes are not executed utile the budget is exhausted or all unlabeled samples are labeled.

V. ASSESSING THE WORKER'S RELIABILITY

The responds generated by different workers are of various qualities on an open crowdsourcing platform, since these non-expert workers have different expertise and experience. Then, we use a small amount of testing samples labeled by experts to assess the workers' reliabilities at first in this work.

Let T be the testing sample set with k labeled samples. For worker w_l, we can get an observed variable set X according to workers' responds and true labels on T. The model parameter Θ describes workers' reliabilities. One of the simplest methods to estimate Θ is based on workers' labeling accuracies on k testing samples. However, we consider the labeled sentences should be treated differently, because each sentence has diverse complexity. Compared with simple sentence, extracting opinion pair(s) from sentence with complex structure or expression is more error-prone even for the reliable worker. If all sentences labeled by a worker are complex, his/her labeling accuracy could be relatively low. Predictably, the worker's reliability would be low, because his/her reliability is estimated based on the labeling accuracy.

Thus, we use a latent variable set Z to model the sentences' complexities. Then, we can estimate the Θ for workers by maximizing the following *Log likelihood* function:

$$\begin{aligned} L(\Theta) = \log P(X|\Theta) &= \log \sum_{Z} P(X, Z|\Theta) \\ &= \log \sum_{Z} P(X|Z, \Theta) P(Z|\Theta) \end{aligned} \quad (2)$$

The Expectation-Maximization (EM) [26] can be used to compute the maximum-likelihood solution above, which consists of two steps: an Expectation (E-step) and a Maximization(M-step).

E-step. Inferring the distribution of latent variable Z according to the $\Theta^{(j)}$ estimated in the j^{th} iteration, and computing the conditional expectation of joint distribution.

$$\begin{aligned} L(\Theta, \Theta^{(j)}) &= \sum_{i=1}^{k} \sum_{C^{(i)}} Q_i(Z^{(i)}) \log P(X^{(i)}, Z^{(i)}|\Theta) \\ &\propto \sum_{i=1}^{k} \sum_{Z^{(i)}} Q_i(Z^{(i)}) \log(P(X^{(i)}|Z^{(i)}, \Theta) P(Z^{(i)}|\Theta)) \end{aligned} \quad (3)$$

where k is the number of testing samples, the auxiliary function $Q_i(Z^{(i)})$ is the distribution function of $Z^{(i)}$:

$$Q_i(Z^{(i)}) = P(Z^{(i)}|X^{(i)}, \Theta^{(j)}) \quad (4)$$

M-step. Estimating $\Theta^{(j+1)}$ by maximizing $L(\Theta, \Theta^{(j)})$:

$$\Theta^{(j+1)} = \arg\max_{\Theta} L(\Theta, \Theta^{(j)}) \quad (5)$$

These two steps (the E- and the M-step) are executed iteratively till the $\Theta^{(j+1)}$ convergence.

VI. ASSIGNING THE LABELING TASKS

Unlike the traditional *one shot*-type task assignment mechanisms, the labeling tasks are assigned to workers based on their reliabilities iteratively in the proposed framework. By this way, we can filter out the unreliable workers to guarantee the quality of responds and reduce the unnecessary labeling cost.

The whole process of harvesting opinion pairs from review sentences with adaptive crowd labeling is described in Algorithm 1, which includes assigning the labeling tasks and integrating the workers' responds. Specifically, we construct a small set T of k testing samples labeled by experts in advance. Each worker's reliability is estimated by this testing sample set with the EM algorithm described above (Line 1 and 2). Based on this assessment, we keep only the relatively high reliable workers in the worker set (Line 3 and 4). Then, a certain proportion of labeling tasks will be assigned to the workers with high reliabilities (Line 7). In order to reduce the labeling cost, a task will be only assigned to several

reliable workers randomly rather than to all reliable ones (Line 8). The responds returned by workers will be used to generate a final result for each task by a respond fusion mechanism (Line 11) presented in next section. Further, some final results with high confidences are put in the result set (Line 12 and 13) based on the their confidence scores, which will be discussed in Section VII. Then, these new generated results are used to reassess the workers' reliability (Line 14). The workers with low reliabilities will be excluded from the worker set again (Line 15), by which we can guarantee the quality of the workers. It is worth noting that we do not need extra cost to label samples for reassessing the workers' reliabilities, because these results would be correct with high probability. The above processes are executed repeatedly till the budget is exhausted or all samples are labeled.

VII. INTEGRATING THE WORKERS' RESPONDS

As discussed above, we will filter the unreliable workers before the labeling tasks are distributed to workers. In order to assure the quality of labeling result set, we assign duplicates of a task to several workers. We will collect multiple responds coming from workers for a sample. Then, we need a mechanism for integrating the responds to generate the final labeling result for each sample.

The simplest method of integrating responds is the majority voting, which treats every vote equally. However, we consider the workers with higher reliabilities should be assigned more authorities on labeling results. On the other hand, the dependency of the opinion target and the corresponding opinion term would provide positive information for integrating the final results.

Given a review sentence s_i, which could include more than one opinion pair. Let V_i be the set of workers labeling sentence i, V_i' denotes the set of workers identifying the j-th opinion pair in s_i. Then, the confidence score SC_i^j of the j-th opinion pair op_i^j in sentence i can be estimated as follows.

$$SC_i^j = D_i^j * \frac{1}{|V_i| - |V_i'| + 1} \sum_{w_k \in V_i'} (1 + \theta_{w_k})^2 \qquad (6)$$

where θ_{w_k} is the reliability of worker w_k, D_i^j is the dependency index of opinion target ot_i^j and opinion term ow_i^j in the j-th opinion pair extracted from sentence s_i by worker w_k.

The dependency index D_i^j is used to quantify the dependency information of opinion target ot_i^j and opinion term ow_i^j. Some potential clues appearing in review sentence would help us to capture the dependency information. For example, the opinion target would probably be noun or noun phrase, opinion target term is likely to be adjective. Then, an opinion pair candidate consisting of a noun/noun phrase and an adjective should has relatively high SC_i^j in Equation 6 under the same conditions. In this work, we capture such dependency information based on six features as follows:

- f_1: The PMI (Pointwise Mutual Information) value of opinion target ot_i^j and opinion term ow_i^j;
- f_2: The frequency of opinion target ot_i^j occurring in review sentence set;
- f_3: The frequency of opinion term ow_i^j occurring in review sentence set;
- f_4: The Part-of-Speech of opinion target ot_i^j;
- f_5: The Part-of-Speech of opinion term ow_i^j;
- f_6: The distance between opinion target ot_i^j and opinion term ow_i^j in sentence s_i.

According to the features above, we construct a vector for each opinion pair candidate labeled by worker. This vector of opinion pair will be used to estimate the similarities with those of existing opinion pair results generated in the previous iterations. That means the opinion pair candidate would be

Algorithm 1 Harvesting Opinion Pairs With Adaptive Crowd Labeling

Input: Worker set W,
Test sample set T,
Unlabeled sample set U,
Budget B,
Percentage of unlabeled samples p,
Duplicate number of an assigned sample d,
Percentage of workers α,
Multiple of confidence score β,

Output: Final result set R;

1: Assigning the testing samples in T to all workers in W for labeling;
2: Assessing the workers' reliabilities by EM process based on their returned responds and ranking these workers by their reliabilities diminishingly;
3: Constructing the set W_h of high reliable workers with the top α percent of workers in the ranking list;
4: $W = W_h$;
5: $B = B - |T|$;
6: **repeat**
7: Generating a set U' by selecting samples in U randomly based on the sampling percentage p;
8: Assigning the samples in U' randomly to workers in W for labeling according to d;
9: $B = B - dp|U|$;
10: $U = U - U'$;
11: Integrating the responds of each sample to generate the final result set R' by a fusion function $f(\bullet)$;
12: Constructing a set R_h by the final results, whose confidence scores are greater than or equal to $\frac{\beta}{|R'|} \sum_{op_i^j \in R'} SC_i^j$;
13: $R = R \bigcup R_h$;
14: Reassessing the reliabilities of workers in W with EM based on R_h and ranking the these workers by their reliabilities degressively;
15: Remaining only the top α percent of workers in the ranking list as the members of W;
16: **until** ($B > 0$ or $U = \varnothing$)
17: **return** R;

likely to be the correct one, if its structure is very similar to that of a true opinion pair in result set R. Let $v_{op_i^j}$ be the corresponding vector of opinion pair op_i^j, r_k be an opinion pair in result set. Thus, we can evaluate the dependency index D_i^j of opinion target ot_i^j and opinion term ow_i^j as follows.

$$D_i^j = \frac{1}{1 + \min\{\bigcup_{r_k \in R}\{sim(v_{op_i^j}, v_{r_k})\}\}} \tag{7}$$

where the *sim* is a distance function for measuring the similarity of two vectors such as Euclidean distance, Cosine similarity.

VIII. EXPERIMENTS

In this work, we target at extracting the pairs of opinion target and opinion term from online reviews. However, such tasks are relatively complicated comparing with traditional crowdsourcing tasks, which makes them cannot be released on existing crowdsourcing platforms. Thus, we have developed a crowdsourcing system for extracting opinion pairs, which includes 273 review sentences from Hu's review dataset [27]. About 72 postgraduate student volunteers serve as the crowdsourcing workers on our platform. Each sentence is labeled five times on average, and 313 opinion pairs are harvested in total eventually.

In order to evaluate the extracting effectiveness of the proposed method, we consider the quantity of correct results generated with certain budget at first. On the other hand, we apply the Precision (P), Recall (R) and F_1-measure (F_1) as follows, which are used commonly for text analysis field [28] and machine learning field [29].

$$P = \frac{|l_a|}{|M|} \quad R = \frac{|l_a|}{|L_a|} \; F_1 = \frac{2PR}{P+R} \tag{8}$$

where M is the set of all generated opinion pairs by integrating the workers' responds, l_a is the set of opinion pairs labeled correctly included in M, L_a is the set of true opinion pairs included in the review set.

We compared our method with a baseline and two state-of-the-art methods based on the evaluation indices above:

1) *MV*. The "Majority Vote" heuristic used commonly for inferring the final results from multiple responds of workers.
2) *GLAD* [25]. A generative model of labels, abilities and difficulties, and some inference methods are used to simultaneously infer worker's expertise, task's difficulty, and the most probable label of each task.
3) *IA* [30]. Tasks are assigned based on a bipartite graph, and an iterative algorithm is used to infer correct answers from workers' responds.

A. RESULTS AND DISCUSSIONS

In the first experiment, we focus on verifying the effectiveness of harvesting the opinion pairs with certain budget. Figure 3 shows that the count of correct opinion pairs by integrating workers' responds with different budgets for four methods, where the ordinate indicates the count of correct opinion pairs generated.

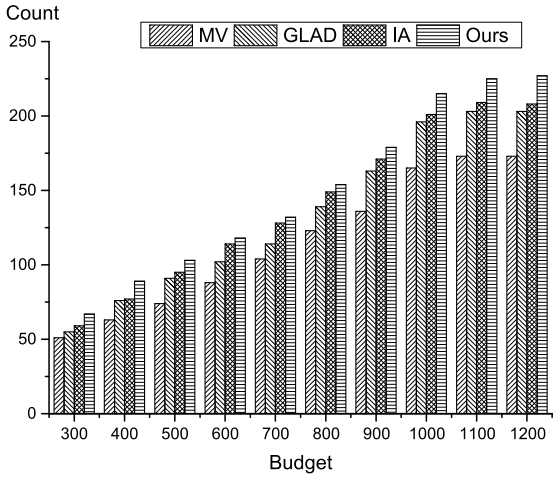


FIGURE 3. The count of correct opinion pairs generated by different methods with certain budget.

As shown in Figure 3, with the increases of budget, four methods' effectiveness on harvesting opinion pairs improve. But the proposed method can always harvest most results with different budgets compared with the competitors. The *MV* performs the worst effectiveness, since it does not consider the difference of workers and treats all responds generated by workers equally. The *GLAD* applies an EM process to infer the final results, which would improve the results' accuracy as the workers' responds increases in quantity. Based on the *GLAD*'s framework, the *IA* uses high reliable responds generated in previous iteration to reestimate worker's reliability and sentence's complexity in each iteration. However, both *GLAD* and *IA* do not filter the low reliable responds, which would lead to a negative impact for the effectiveness.

Next, we compare the precisions, recalls, F1-measures and AUC of our method with the competitors' according to different budget constraints in Table 1. Notably, we apply the precision-recall curve rather than the receiver operating characteristic curve for the AUC. Because the opinion pair extraction is essentially a classification task with skewed data, and the precision-recall curve can give a more accurate picture of an algorithm's performance for such tasks [31]. As shown in Table 1, we can find that the effectiveness of each method is improved with budget rising. But our method achieves the best effect under various budget constraints for almost all evaluation indices.

Since worker reliability and dependency information are main component of Equation 6, we investigate their respective contributions to the confidence score of opinion pair in the following experiment. As shown in Table 2, we can find that both worker reliability and dependency information make positive effect for acquiring the opinion pairs. However, the worker reliability is more effective than the dependency information. We consider that is because a reliable worker would use the sentence structure unconsciously to determine

TABLE 1. The comparisons on Precision, Recall, F_1-measures and AUC of different methods with various budget.

Budget		300	400	500	600	700	800	900	1000	1100	1200
MV	P	0.630	0.594	0.565	0.564	0.575	0.597	0.574	0.573	0.560	0.560
	R	0.163	0.201	0.236	0.281	0.332	0.393	0.435	0.527	0.553	0.553
	F_1	0.259	0.301	0.333	0.375	0.421	0.474	0.495	0.549	0.556	0.556
	AUC	0.754	0.743	0.737	0.735	0.737	0.744	0.741	0.742	0.741	0.741
GLAD	P	0.679	0.717	0.695	0.654	0.630	0.680	0.688	0.681	0.657	0.657
	R	0.176	0.243	0.291	0.326	0.364	0.447	0.521	0.626	0.649	0.649
	F_1	0.279	0.363	0.410	0.435	0.462	0.539	0.593	0.652	0.653	0.653
	AUC	0.813	0.821	0.810	0.791	0.785	0.805	0.807	0.811	0.810	0.810
IA	P	0.728	0.726	0.725	**0.731**	**0.707**	**0.728**	0.722	0.698	0.673	0.673
	R	0.188	0.246	0.304	0.364	0.409	0.479	0.546	0.642	0.665	0.665
	F_1	0.299	0.368	0.428	0.486	0.518	0.578	0.622	0.669	0.669	0.669
	AUC	0.819	0.815	0.797	0.805	0.800	0.822	0.827	0.829	0.824	0.824
Ours	P	**0.801**	**0.782**	**0.746**	0.714	0.693	0.711	**0.732**	0.739	**0.713**	**0.714**
	R	**0.212**	**0.289**	**0.324**	**0.373**	**0.416**	**0.493**	**0.582**	**0.701**	**0.727**	**0.727**
	F_1	**0.335**	**0.422**	**0.452**	**0.490**	**0.520**	**0.582**	**0.648**	**0.719**	**0.720**	**0.721**
	AUC	**0.869**	**0.867**	**0.846**	**0.833**	**0.825**	**0.840**	**0.839**	**0.839**	**0.836**	**0.837**

TABLE 2. The contributions of worker reliability and dependency information to extraction results.

	Budget									
	300	400	500	600	700	800	900	1000	1100	1200
worker reliability	67	90	101	116	130	154	179	215	225	227
dependency information	9	11	18	20	24	31	35	39	39	39
worker reliability+dependency information	66	90	104	122	137	161	220	220	228	228

the opinion pairs. When the budget is 300, the dependency information seems to have a negative influence for extraction effectiveness. The main reason is that small budget would lead to small result set and the dependency index is calculated based on this result set by a distance function. Then, the dependency information of opinion target and opinion term cannot be acquired correctly. We also find that the dependency information acts progressively more with increasing budget in Table 2. That is mainly because more and more results would be generated with the increase in budget.

In the proposed method, we estimate the worker's reliability based on a prepared testing set at first. The initial testing set includes 22 opinion pairs extracted from 18 review sentences in our experiment. In each subsequent iteration, some integrated results with high confidences are used to enlarge the testing set for reassessing the workers' reliabilities without extra cost. Thus, we investigate the influences of different counts of additional testing samples on the proposed method's effectiveness in the third experiment. Then, we enlarge the testing set with different proportions of results generated in previous iteration, which is described as the abscissa in Figure 4. We can find that if more results are used to enlarge the testing set, better performance will be achieved for each budget. That means that sufficient testing samples would ensure the evaluation accuracy of workers' reliabilities. It is worth noting that we do not need extra cost for this reassessment process, because we only use the generated results to enlarge the testing set.

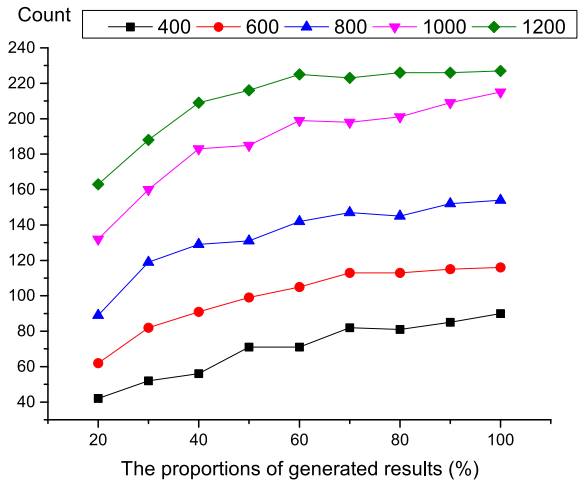


FIGURE 4. The impact on the quantity of testing samples.

Assigning a task to multiple workers for labeling is an effective way to ensure the quality of generated results, which is adopted in most existing methods. We compare the proposed method with the baselines according to different counts of task duplicates under the budget 1100 in

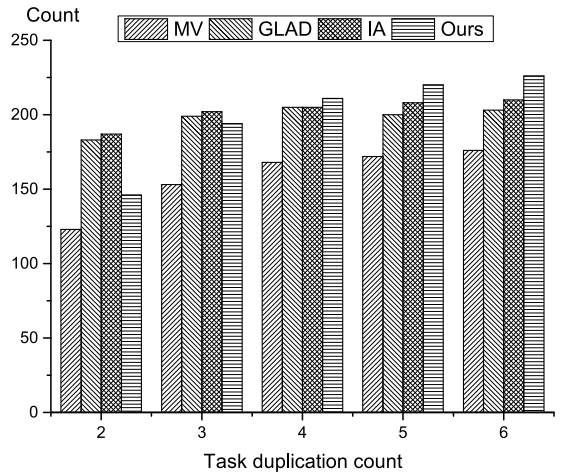


FIGURE 5. Task assignment with different count of duplicates.

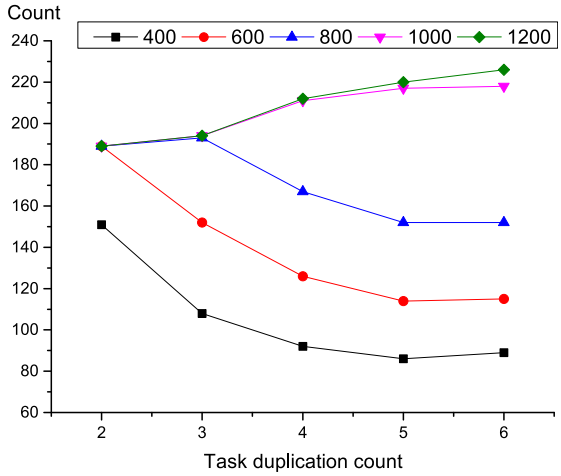
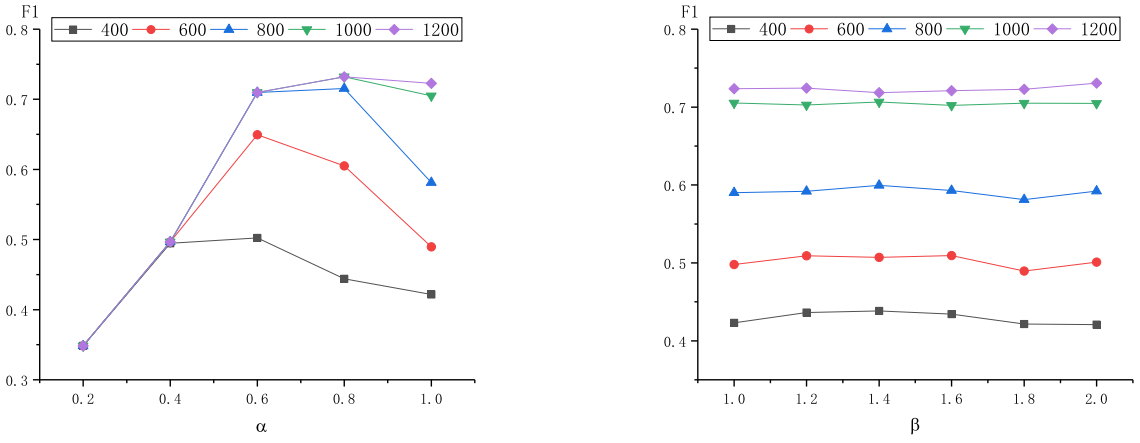


FIGURE 6. The effect of task duplicates.

Figure 5. As the duplicate count increases, the counts of harvested opinion pairs increase for each method. However, the proposed method achieves the best performance when the duplicate count exceeds 3. Furthermore, we can find that the *GLAD* and *IA* are not affected appreciably by the duplicate count. It is because these two methods extract the opinion pairs by making a global analysis on all responds, and the proposed method treats the review sentence separately. Thus, the proposed method can achieve better effectiveness when the task duplicates assigned are sufficient.

We further investigate the effect of task duplicates for the proposed method in the following experiment. We compare the counts of harvested opinion pairs according to different duplicate counts under various budget constraints in Figure 6. When the budget is relatively small (less than or equal to 600 in our experiment), the harvesting effectiveness declines as duplicate count raises. However, if the budget increases to certain amount (greater than or equal to 1000 in our experiment), the harvesting effectiveness will improve significantly as duplicate count raises. We think that is because the sufficient budget can collect more responds from workers, which can be used to generate the final results. On the other hand, the generated results will improve the quality of testing set further. Thus, we can estimate the workers' reliabilities more accurately, and then more opinion pairs will be harvested.

At the last experiment, we investigate the influences of parameters α and β in Algorithm 1 on harvesting the opinion pairs under various budget constraints. The former is used to determine which workers are treated as the reliable ones in the worker set. The latter is used to determine which results are those with high confidences. Figure 7 shows the influences of these two parameters on harvesting the opinion pairs. We can find that 80% seems to be a good choice for α in the cases of relatively sufficient budget in the first subgraph. When the budget is limited, we should set α to about 60%. For the parameter β, it seems to have no influence on harvesting effectiveness apparently, since different values of β achieve the similar F_1 score. However, the precisions and recalls vary greatly for different values of β. For example, the precision is 73.5% and the recall is 70.9% when β is set to 2 and the



(a) The effect of α

(b) The effect of β

FIGURE 7. The influences of parameters α and β on harvesting the opinion pairs.

budget is 1100. But the precision drops to 66.6% and the recall increases to 77.6% for the same budget, when β is set to 1. Then, we should set the value of β according to the specific requirement in practical applications.

IX. CONCLUSION

Online reviews play important roles in many Web Applications like e-business and government intelligence. The pair of opinion target and opinion term is vital component for expressing the user's opinion. The supervised methods can work well for extracting the opinion pairs from reviews. In this scenario, it is essential to construct a set of labeled samples with high quality for training the extractor model. However, constructing such a training set by traditional manual labeling way is laborious, error-prone and cost-consuming.

In this work, we explore the problem of constructing the set of opinion pairs iteratively from reviews by crowdsourcing service under a budget constraint. Specifically, we propose a task assignment mechanism based on a forward assessment process on worker's reliability. By this way, we can guarantee the quality of workers' responds by filtering the unreliable workers. And then, we integrate multiple workers' responds of a task into a final result by considering workers' reliabilities as well as the dependence relation between opinion target and opinion term. In order to assure the workers' reliabilities, we use the results generated in previous iteration to reassess the workers without extra cost. The experimental results show that the proposed method can achieve better extraction effectiveness compared with the traditional methods and the state-of-the-art ones.

REFERENCES

[1] H. Wang, C. Zhang, H. Yin, W. Wang, J. Zhang, and F. Xu, "A unified framework for fine-grained opinion mining from online reviews," in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 1134–1143.

[2] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers*, Dec. 2016, pp. 3298–3307.

[3] Y. Lin, X. Jiang, Y. Li, J. Zhang, and G. Cai, "Semi-supervised collective extraction of opinion target and opinion word from online reviews based on active labeling," *J. Intell. Fuzzy Syst.*, vol. 33, pp. 3949–3958, Jan. 2017.

[4] V. C. Raykar, S. Yu, and L. H. Zhao, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Jun. 2009, pp. 889–896.

[5] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based Bayesian aggregation models for crowdsourcing," in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 155–164.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[7] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 2334–2346.

[8] T. Mitra, C. J. Hutto, and E. Gilbert, "Comparing person-and process-centric strategies for obtaining quality data on Amazon mechanical Turk," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 1345–1354.

[9] J. Feng, G. Li, and J. Feng, "Research summarization on crowdsourcing technology," *Chin. J. Comput.*, vol. 38, no. 9, pp. 1713–1726, 2015.

[10] H. Jiang and S. Matsubara, "Efficient task decomposition in crowdsourcing," in *Proc. PRIMA*, Dec. 2014, pp. 65–73.

[11] M. Brambilla, S. Ceri, A. Mauri, and R. Volonterio, "An explorative approach for crowdsourcing tasks design," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 1125–1130.

[12] C. J. Zhang, L. Chen, and Y. Tong, "MaC: A probabilistic framework for query answering with machine-crowd collaboration," in *Proc. CIKM*, Nov. 2014, pp. 11–20.

[13] Z. Guo, C. Tang, W. Niu, Y. Fu, T. Wu, H. Xia, and H. Tang, "Fine-grained recommendation mechanism to curb astroturfing in crowdsourcing systems," *IEEE Access*, vol. 5, pp. 15529–15541, 2017.

[14] M. Z. Tunio, H. Luo, W. Cong, Z. Fang, A. R. Gilal, A. Abro, and S. Wenhua, "Impact of personality on task selection in crowdsourcing software development: A sorting approach," *IEEE Access*, vol. 5, pp. 18287–18294, 2017.

[15] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—But is it good?: Evaluating non-expert annotations for natural language tasks," in *Proc. EMNLP*, Oct. 2008, pp. 254–263.

[16] X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang, "CDAS: A crowdsourcing data analytics system," in *Proc. VLDB Endowment*, Jun. 2012, pp. 1040–1051.

[17] W. Li, C. Zhang, Z. Liu, and Y. Tanaka, "Incentive mechanism design for crowdsourcing-based indoor localization," *IEEE Access*, vol. 6, pp. 54042–54051, 2018.

[18] C. Li, V. S. Sheng, L. Jiang, and H. Li, "Noise filtering to improve data and model quality for crowdsourcing," *Knowl.-Based Syst.*, vol. 107, pp. 96–103, Sep. 2016.

[19] C. Li, L. Jiang, and W. Xu, "Noise correction to improve data and model quality for crowdsourcing," *Eng. Appl. Artif. Intell.*, vol. 82, pp. 184–191, Jun. 2019.

[20] P. Donmez, J. G. Carbonell, and J. Schneider, "Efficiently learning the accuracy of labeling sources for selective sampling," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jun. 2009, pp. 259–268.

[21] H. J. Jung and M. Lease, "Inferring missing relevance judgments from crowd workers via probabilistic matrix factorization," in *Proc. SIGIR*, Aug. 2012, pp. 1095–1096.

[22] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 614–622.

[23] B. Jia, H. Xu, S. Liu, and W. Li, "A high quality task assignment mechanism in vehicle-based crowdsourcing using predictable mobility based on Markov," *IEEE Access*, vol. 6, pp. 64920–64926, 2018.

[24] C. Qiu, L. Jiang, and Z. Cai, "Using differential evolution to estimate labeler quality for crowdsourcing," in *Proc. PRICAI*, Jul. 2018, pp. 165–173.

[25] J. Whitehill, T. F. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2035–2043.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[27] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proc. AAAI*, Jul. 2004, pp. 755–760.

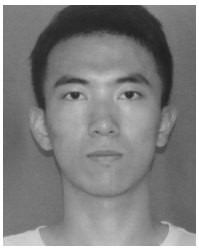
[28] L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive Bayes and its application to text classification," *Eng. Appl. Artif. Intell.*, vol. 52, pp. 26–39, Jun. 2016.

[29] L. Jiang, L. Zhang, C. Li, and J. Wu, "A correlation-based feature weighting filter for naive Bayes," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 2, pp. 201–213, Feb. 2019.

[30] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Proc. NIPS*, 2011, pp. 1953–1961.

[31] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proc. ICML*, Jun. 2006, pp. 233–240.

YUMING LIN was born in 1978. He received the B.S. and M.S. degrees in computer science and technology from the Guilin University of Electronic Technology, Guilin, China, and the Ph.D. degree in computer application from the East China Normal University, Shanghai, China. He is currently an Associate Professor of computer sciences with the Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology. His current research interests include opinion mining, knowledge graph, and massive data management.



WEI ZHAO was born in 1995. He received the B.S. degree from the Zhengzhou University of Light Industry, Zhengzhou, China. He is currently pursuing the M.S. degree with the School of Computer Science and Information Security, Guilin University of Electronic Technology and acting as an intern with the School of Data Science and Engineering, East China Normal University. His current research interest includes opinion mining.

YOU LI was born in 1980. She received the B.S. degree in computer software from the Guilin University of Electronic Technology, China, and the M.S. degree in computer science from the Dalian University of Technology, Dalian, China. She is currently an Associate Professor of computer sciences with the Guangxi Key Laboratory of Automatic Detecting Technology and Instruments, Guilin University of Electronic Technology. Her current research interests include natural language processing and machine learning.

HUIBING ZHANG was born in 1976. He received the B.S and M.S degrees in computer science and technology from the Guilin University of Electronic Technology, Guilin, China, in 2007, and the Ph.D. degree in computer science and technology from the Beijing University of Technology, Beijing, China, in 2012. He is currently an Associate Professor with the Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology. His research interests include trust evaluation and management in the Internet of Things, and social computing.

YA ZHOU was born in 1966. She received the M.S degree in computer science from Fudan University, China. She is currently a Professor with the Guilin University of Electronic Technology. Her research interests include distributed systems, database theory, data mining, and web service technology.