Applied Intelligence (2022) 52:9101–9116
https://doi.org/10.1007/s10489-021-02958-3

Multiple weak supervision for short text classification

Li-Ming Chen[1] · **Bao-Xin Xiu**[2] · **Zhao-Yun Ding**[1]

Accepted: 26 October 2021 / Published online: 1 January 2022
© The Author(s) 2021

Abstract
For short text classification, insufficient labeled data, data sparsity, and imbalanced classification have become three major challenges. For this, we proposed multiple weak supervision, which can label unlabeled data automatically. Different from prior work, the proposed method can generate probabilistic labels through conditional independent model. What's more, experiments were conducted to verify the effectiveness of multiple weak supervision. According to experimental results on public dadasets, real datasets and synthetic datasets, unlabeled imbalanced short text classification problem can be solved effectively by multiple weak supervision. Notably, without reducing *precision*, *recall*, and *F1-score* can be improved by adding distant supervision clustering, which can be used to meet different application needs.

Keywords Multiple weak supervision · Short text classification · Imbalanced classification · Distant supervision clustering · Probabilistic labels

1 Introduction

Traditionally, supervised machine learning relies on useful feature representation and hand-labeled data. With deep learning techniques, useful feature representation can be learned easily [1]. However, for supervised machine learning, deep learning cannot function without sufficient labeled data [2]. Moreover, the requirements for data labels usually evolve rapidly as applications change [3]. These changes can be labeling guidelines, labeling granularity [4], application scenarios and so on. What's more, most training data samples are still labeled manually, which may be extremely expensive and time-consuming [3]. Thus, there is an urgent need for an efficient method to label training data automatically, especially for short text classification. Secondly, data sparsity remains a key challenge for short text classification [4]. Thirdly, in the real world, text classification is usually imbalanced. That is, short text classification is usually faced with insufficient labeled data, data sparsity and imbalanced classification simultaneously.

To address insufficient labeled data, data sparsity and imbalanced classification in short text classification wholly, multiple weak supervision [1, 5] was proposed, where conditional independent model was introduced to generate probabilistic labels as accurate as possible. To be specific, to label short text data automatically, three kinds of weak supervision sources (*keywords matching*, *regular expressions* and *distant supervision clustering*) were creatively introduced. Notably, *keywords matching* and *regular expressions* were used to represent explicit knowledge, while *distant supervision clustering* was specially designed to represent tacit knowledge.

Specially, *distant supervision clustering* was proposed firstly in this paper. According to the process, distant supervision clustering can be divided into three steps. The first step is to specify the similarity threshold, which is the criteria of distant supervision clustering. The second step is to calculate the similarity between the sample points and knowledge base. The third step is to compare the calculated similarity with the similarity threshold. If the calculated similarity is no less than the similarity threshold, the sample point will be labeled as the same as the corpus. Otherwise,

✉ Li-Ming Chen
chenliming18@nudt.edu.cn

Bao-Xin Xiu
baoxinxiu@163.com

Zhao-Yun Ding
zyding@nudt.edu.cn

[1] Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, China

[2] School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China

the sample point will get *abstain* label. In fact, similarity threshold plays a key role in distant supervised clustering. However, since this paper focuses on the proposal of multiple weak supervision framework, similarity threshold will not be studied in depth. For example, the impact of similarity threshold and other related studies will be reflected in the follow-up work. Please look forward to it.

2 Related work

According to the first chapter, short text classification in real scenes is usually faced with three major challenges: insufficient labeled data, data sparsity and imbalanced classification. There are few comprehensive studies on labels bottleneck (insufficient labeled data), data sparsity and imbalanced classification. In fact, existing research usually focuses on solving only one problem. Therefore, the related work of labels bottleneck, data sparsity and imbalanced classification will be introduced one by one as follows.

2.1 Labels bottleneck

To solve labels bottleneck (insufficient labeled data) in natural language processing, there are two main solutions: weak supervision as well as *fine tuning*. Weak supervision is committed to expanding the scale of labeled data from the data level. Differently *fine tuning* aims to provide an initialization model as good as possible, so as to reduce the requirement of labeled data scale.

(1) Weak Supervision There are many attempts to label training data in programmatic way. Generally speaking, these labeling ways, which generate nosier weak labels based on domain knowledge [1], are referred to as weak supervision. Taking text classification for example, if a text contains any one of certain keywords, it can be classified as a specific category. Distant supervision, the most popular one, gets weak labels by aligning the data points with the external knowledge base [7–9]. In addition, crowdsourcing labels [10, 11], heuristic rules for labeling data [12, 13] and others [14–17] are also the common sources of weak supervision. That is, weak supervision sources mainly contain distant supervision [18–20], crowdsourcing [10, 11] and heuristic rules [12, 13].

Distant supervision is mainly used for relation extraction [8, 21, 22]. The main idea is to align the sample points with the records in the external database [19]. For example, distant supervision can be used to extract spouse relation by aligning the sample points with the spouse records of an external knowledge base [1] (such as DBpedia [23] and Wikipedia [22]). Obviously, the external knowledge base needs to have a relative strong correlation with the target task. However, such a highly relevant knowledge base is usually scarce, which hinders the extended application of distant supervision.

Crowdsourcing, also called human computation [11, 24], is the process that a number of non-experts collectively perform a labeling task [25]. The explosive growth and widespread accessibility of the Internet have led to the surge of crowdsourcing [11]. Crowdsourcing has been widely used in labeling tasks of machine learning, which require a lot of human computation but little domain knowledge. These areas include image and video annotation [26–28], named entity annotation [11], relevance evaluation [29], natural language annotation [30–32] and others [11, 33]. Crowdsourcing can quickly generate a large number of data labels, but the quality of data labels is relatively poor.

Heuristic rules for labeling data are usually written by users or domain experts [3]. Due to the diverse quality of heuristic rules, the accuracy and correlation of labels might fluctuate widely [13, 34]. Therefore, the efficiency of rules-based labeling strategy depends on the quality of heuristic rules [35]. In view of this, the heuristic rules (or domain knowledge) from domain experts are essential for high quality labels.

However, any kind of weak supervision is weak and limited. This is because a kind of weak supervision is no longer sufficient to generate large higher-quality data labels. In light of this, to alleviate labels shortage, multiple weak supervision were introduced for labeling short text data. To be specific, according to the characteristics of short text classification, we combine keywords matching, regular expressions and distant supervision clustering to label short text and train classifier.

(2) Fine Tuning In natural language processing, inadequate labeled data is usually too less to be used to learn good enough model parameters. Based on this, *fine tuning* were proposed to reduce the amount of labeled data needed for parameter learning. In short, the pre-training model can provide a good parameter initialization for tasks with insufficient labeled data. Based on this good parameter initialization, the model training only needs to fine tune the parameters to achieve the optimal solution. For this, fine tuning is usually done with a small amount of labeled data.

In conclusion, the pre-training model directly determines the quality of parameter initialization. At present, the pre-training models for text mainly included ELMo [36], GPT (Generative Pre-Training) [37], BERT (Bidirectional Encoder Representation from Transformers) [38], XLNet [39], ZEN [40], ERNIE (Enhanced Language Representation with Generative Entities) [41], etc. In particular, BERT and ERNIE have attracted a lot of attention and derived

some deformation, such as RoBERTa [42] and ERNIE2.0 [43].

However, the training of the pre-training model requires a large amount of computing resources. For example, the training of BERT model [39], in the Google 64 TPU computing environment, still lasted for nearly 4 days. In addition, as time goes on, fixed pre-training models are prone to problems such as "concept drift" and even lack of generalization ability. Last but not least, fine tuning relies on strong labeled data, which cannot be provided by weak supervision. Therefore, the pre-training model is not very suitable for short text classification with multipe weak supervision.

2.2 Data sparsity

With the growth of instant messaging by Mobile Internet, the proliferation of short texts highlights the challenge of data sparsity and misspelling (informal writing) [54, 58], which limits the application of machine learning in short text classification. To address these problem, two types of solutions were proposed: feature strategy and algorithm strategy (Table 1 [46]). Notably, in feature selection, the measure of filter-based approach can be chi-squared (CHI2) [76], information gain (IG) [77, 78], correlation coefficient (CC) [79], accuracy balanced (Acc2) [80], pointwise mutual information (PMI) [61, 81], odds ratio (OR) [82] and multi-class odds ratio (MOR) [69].

Undoubtedly, both feature strategy and algorithm strategy have good effect on supervised learning with large-scale labeled data. However, they all did not considevr the case of data sparsity with weak supervision learning. Moreoverv, even data augmentation cannot really address the simultaneous challenges of insufficient lableled data, data sparsity, and imbalanced classification. In particular, data augmentation will also bring uncontrollable semantic changes, which will further increase the challenge of clasification. Similarly, distributed representations, such as word2vec and Glove, are difficult to be directly incorporated into multiple weak supervision framework due to their high computational overhead and dependence on strong labeled data sets.

For the sake of simplicity, only *N-gram* is taken as an example to carry out experimental test. In light of this, for short text classification with weak supervision, *N-gram* (feature representation) and *Logistic Regression* (algorithm) were introduced for addressing data sparsity and misspelling. Taking one step further, to solve data sparsity, *N-gram* (feature representation) and *Logistic Regression* (algorithm) were embedded into the proposed multiple weak supervision framework. Such a design is for simplicity and practicality. As for the dimension disaster that N-gram may cause, this paper does not rule out. The related ablation

Table 1 Solution of data sparsity [46]

Solution/Strategy	Process/Step	Approach/Algorithm
Feature Strategy	Feature Representation	*bag-of-words* [44, 45]
		N-grams [46, 47]
		TF (Term Frequency) [46, 48]
		TF-IDF [48–50]
	Feature Selection	*Filter-based approach* [51–53]
		Wrapper Approach [54, 55]
		Embedded Approach [56]
		Hybrid Approach [51]
	Feature Extraction	*Partial least square* [57]
		Latent semantic indexing [58]
		PCA (Principal component analysis) [59, 60]
Algorithm Strategy	Single Algorithm	*NB (Naïve Bayes)* [61, 62]
		DT (Decision Tree) [63]
		KNN (K-Nearest Neighbor) [64, 65]
		SVM (Support Vector Machine) [44, 51, 66]
		NN (Neural Networks) [49]
		Logistic regression [44, 46]
	Ensemble Algorithm	*Random forest* [44, 48]
		Bagging [46, 67, 68]
		Dagging [46, 67, 68]
		Boosting [46, 67, 68]

research will be further carried out in the following research. After all, this article focuses more on proposing a solution framework to solve the classification of unlabeled short texts.

2.3 Imbalanced classification

Imbalanced classification is a hotspot in data mining, machine learning and pattern recognition. There are several top-level conferences devoted to discussing and studying imbalanced classification problem, such as ICML 2003 [70], ACM SIGKDD2004 [71] and IJCAI 2017 [72]. In short, there are mainly four factors influencing the imbalanced classification problem: 1) the scale of the training set; 2) category priority; 3) the misclassification costs of different categories; 4) the location of the boundary.

In general, imbalanced classification has two major research directions: data strategy and algorithm strategy. By changing the distribution of original dataset, the data strategy increases the minority samples (over-sampling) [73–75] or decreases the majority samples (under-sampling) [76–78], so that the imbalanced data tends to balance. This strategy is favored by many researchers because of its advantages in improving the classification performance and being suitable for various classifiers [79]. Although there are more studies on over-sampling than under-sampling, it is still difficult to give a conclusion that over-sampling is better than under-sampling. Therefore, some studies also put forward the mixed sampling method, that is, the method of balancing the training set by synthesizing over-sampling and under-sampling [80].

By contrast, the algorithm strategy mainly makes the classification more focused on minority classes by means of weighting, voting, iteration and so on. Specifically, common methods include cost-sensitive learning and ensemble learning. Cost-sensitive learning was put forward to focus on imbalanced classification of minority classes. It mainly increases the misclassification cost of minority classes with cost-sensitive factor [81]. That is, learning parameters are adjusted to highlight the importance of minority classes. Theses parameters mainly have data space weighting, cost matrix of category dependence, and ROC (receiver operating characteristic curve) threshold. In addition, ensemble learning is also favored [82]. The basic idea [67, 83] is to train a series of basic classifiers and then improve the classification accuracy through integration. Bagging, Boosting and Random Forest are the most commonly used ensemble methods. There are two main reasons why research on algorithm strategy is less than that on data strategy. First, the determination of the cost matrix is very difficult; second, the cost sensitivity depends on different classifiers [81, 84]. As a result, researchers tend to integrate the algorithm strategy into the classification research of specific background rather than as a single research point. But the algorithm strategy is difficult to popularize, whose promotion application cost is very high. Based on this, a resolution mechanism, which is based on probabilistic labels generated from conditional independent model, was put forward to handle imbalanced classification.

For one thing, data strategy is easy to destroy the original distribution and requires very proper sampling methods. For another thing, algorithm strategy is hard to popularize and has very high promotion application cost. Motivated by this, a resolution mechanism based on probabilistic labels generated from conditional independent model, was put forward to handle imbalanced classification of weak supervision.

To sum up, any one of existing methods is hard to address labels shortage, data sparsity and imbalanced classification simultaneously. In other words, there is hardly effective overall solutions for the tree challenges. In light of this, an overall methodology, which is on the basis of multiple weak supervision and probabilistic labels, was proposed and elaborated in chapter 5.

3 Domain knowledge in weak supervision

In order to select proper weak supervision combination, dynamic theory was chosen as the guidance [85]. According to [85], domain knowledge can be divided into explicit knowledge and tacit knowledge. Corresponding to weak supervision, the relation between domain knowledge and weak supervision sources was shown in Fig. 1.

As shown in Fig. 1, explicit knowledge can be represented by heuristic rules, while tacit knowledge involves distant supervision and crowdsourcing labels. In spired by this, to combine both explicit knowledge and tacit knowledge [85], we adopt three types of weak supervision sources: simple keywords matching, regular expressions, and distant supervision clustering. Correspondingly, these three types can be boiled down to two categories: heuristic rules and distant supervision clustering, which correspond to explicit knowledge and tacit knowledge respectively.

3.1 Explicit knowledge (heuristic rules)

In order to represent explicit knowledge, two types of heuristic rules were designed to label data automatically. Specifically, simple keywords matching as well as regular expressions were adopted as explicit knowledge sources.

Combining keyword matching with regular expressions, nearly all explicit knowledge for text classification can be represented easily. However, tacit knowledge is hard to represent. Furthermore, it is prohibitively hard to get high *recall* score with the limited coverage of heuristic rules. In

view of this, distant supervision clustering was proposed to represent tacit knowledge and improve *recall* score.

3.2 Tacit knowledge (distant supervision clustering)

As shown in Table 2, explicit knowledge, hard to quantify, can be represented formally by heuristic rules. On the contrary, tacit knowledge is easy to quantify while it is difficult to represent explicitly. In view of this, distant supervision clustering, a novel weak supervision strategy, was proposed to represent explicit knowledge.

Notably, distant supervision clustering was inspired by distant supervision. For one thing, distant supervision, as a popular weak supervision source, can be regarded as one of the semi-supervised learning methods. Instead of the alignment strategy of distant supervision, distant supervision clustering gets weak labels based on cluster assumption. To be specific, the implication of the *cluster assumption* is that the data has a cluster structure and that the same cluster sample belongs to the same category. This is consistent with the clustering hypothesis of semi-supervised learning [4, 6].

Algorithm 1 Distant supervision clustering.

Input: unlabeled dataset $Train = (x_1), (x_2)..., (x_N)$; Small-scale labeled dataset$Dev = (x_1, y_1), (x_2, y_2)..., (x_l, y_l)$; Similar dataset $KB = \{(x_1, x_2..., x_M), Y_{kb}\}$;
Output: weak label vector L_{ij}.
Step1 determine the similarity threshold
1 **for** $i = 1$ **to** l **do**
2 **for** $j = 1$ **to** M **do**
3 **if** $y_i == Y_{kb}$ **then**
4 $S_{ij} \leftarrow cos(x_i, KB_j)$
5 Similarity threshold $\alpha \leftarrow max(S_{ij})$
Step2 calculate the similarity
1 **for** $i = 1$ **to** N **do**
2 **for** $j = 1$ **to**N_{kb} **do**
3 Similarity $S_i \leftarrow maxcos(x_i, KB_j)$
Step3 Data Labeling
1 **for** $i = 1$ **to** N **do**
2 **if** $S_i > \alpha$ **then**
3 $L_i \leftarrow Y_{kb}$
4 **else**
5 $L_{ij} \leftarrow Abstain$

As shown in Algorithm 1, distant supervision clustering can be divided into 3 steps *Determining Threshold*, *Calculating Similarity*, and *Assigning Labels*.

Specially, the similarity threshold is the maximum similarity between the small-scale labeled dataset and the external corpus. It is noted that similarity threshold, plays an important role in distant supervision clustering and tacit knowledge representation. For one thing, a proper threshold could ensure the quality (accuracy) of labels from distant supervision clustering. For another thing, if threshold is small enough, the vast majority of samples in the corresponding category will receive labels from it, which means very high *recall* score. Most importantly, with distant supervision clustering, we can represent tacit knowledge by quantitative method, which is hard to be represented formally by heuristic rules. However, since this paper focuses on the proposal of weak supervision framework, it will not be studied in depth. For example, the impact of similarity threshold and other related studies will be reflected in the follow-up work. Please look forward to it.

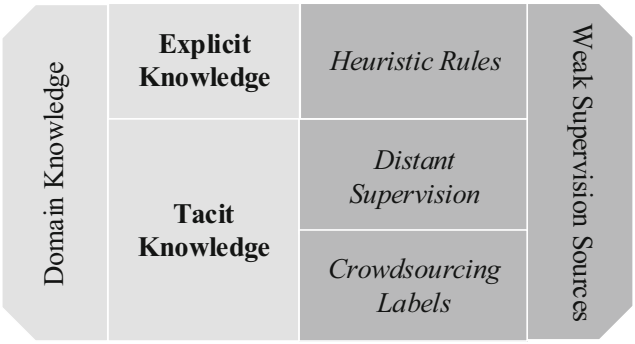

Fig. 1 Domain knowledge represented by weak supervision

In this way, explicit knowledge can be represented by heuristic rules (simple keywords matching and regular expressions) while tacit knowledge can be included in distant supervision clustering. Thus, the coverage and quality of weak labels of training data can be obtained and applied to short text classification by machine learning. In next chapter, the labels integration mechanism and probabilistic labels suitable for solving imbalanced problem will be introduced in detail.

Specially, LDA (Latent Dirichlet Allocation) [86] was bringing in for extracting explicit knowledge (keywords pattern) extraction. LDA is a generative probabilistic model of a corpus. In LDA (Fig. 2), documents are represented as random mixtures over latent topics while each topic is characterized by a distribution over words. Dirichlet Allocation was thought to be the prior distribution of parameter of topic distribution. Notably, compared with common TF-IDF and TextRank model [87], LDA is more suitable for short text classification. Moreover, LDA can also better meet the background constraints, such as data

Table 2 Representation difference between explicit and tacit knowledge

Domain knowledge	Formal representation	Quantifiable
Explicit Knowledge	*Easy*	*No*
Tacit Knowledge	*Hard*	*Yes*

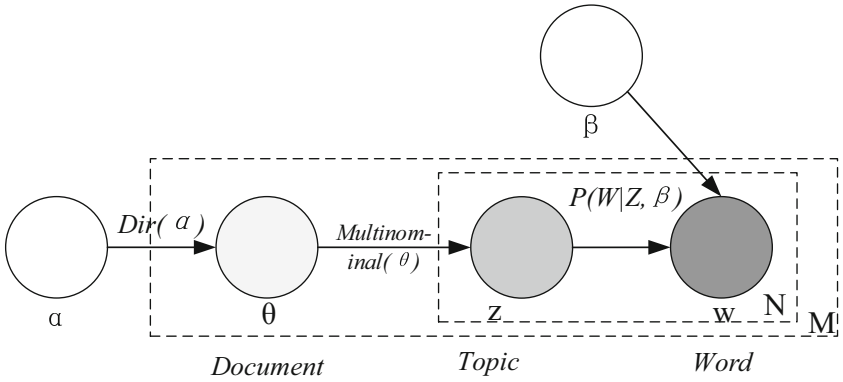

Fig. 2 Graphical Model of LDA (Latent Dirichlet Allocation)

sparsity, simplicity and limited space. Additionally, in the case of multiple weak supervision, the performance comparison of different keyword extraction algorithms will be elaborated in the following research and papers.

Taking binary classification for example, with LDA and prior (explicit) knowledge, we can get some keywords closely related to positive and negative class. With these keywords, we can quickly classify some data points to a category. To be specific, small-scale labeled dataset (e.g. *Dev* in Chapter 6) can be used to build LDA model and extract keywords of specific class. Despite of this, single keywords matching is not always useful for the flexibility and diversity of natural language expressions. Thus, regular expressions were absorbed to accommodate more complex expressions. For example, *"check*out"* can match any character other than the newline character 0 or more times between "check" and "out".

4 Probabilistic labels for imbalanced calssification

With traditional method, data label y_i of binary classification is usually in following format:

$y_i \in Y = \{-1, +1\}, i = 1, 2, \cdots, n;$

where -1 and 1 correspond to negative class and positive class respectively. Based on this, y_i can also be formally represented as labels matrix $L_{n\times 2}$: $L_{n\times 2} =$

$$\begin{bmatrix} y_{11}, & y_{12} \\ y_{21}, & y_{22} \\ \cdots, & \cdots \\ y_{n1}, & y_{n2} \end{bmatrix} \tag{1}$$

where each row i corresponds to one piece of data, and each column j corresponds to a category; $y_{ij} \in Y' = \{0, 1\}$; 0,1 indicate whether they belong to the corresponding category or not; each row has only one value of 1. More generally, the k-classification ($k \geq 2, k \in Z$) problem is as $n \times k$ matrix $Ł_{n\times k} =$:

$$\begin{bmatrix} y_{11}, & y_{12}, & \cdots, & y_{1k} \\ y_{21}, & y_{22}, & \cdots, & y_{2k} \\ \cdots, & \cdots, & y_{ij}, & \cdots \\ y_{n1}, & y_{n2}, & \cdots, & y_{nk} \end{bmatrix} \tag{2}$$

where each row i corresponds to one piece of data, and each column j corresponds to a category; $y_{ij} \epsilon Y' = \{0, 1\}$; 0,1 indicate whether they belong to the corresponding category or not; each row has only one value of 1.

Even though labels matrix $Ł_{n\times k}$ has $n \times k$ elements, there are only n non-zero elements. In fact, the sparsity of labels matrix is rooted in the "black or white" indicator of discrete labels. By contrast, labels of weak supervision tend to be gray, or probabilistic, rather than discrete. Therefore, compared with discrete labels, probabilistic labels are more suitable for representing labels from weak supervision. According to [76], imbalanced classification refers to different sample sizes of different categories. Specifically, the category here refers to the discrete labels. Taking one step further, imbalanced classification is named because of the imbalance distribution of discrete labels among different categories. That is, imbalanced classification may be alleviated by replacing discrete labels with probabilistic labels. For illustration, let's take the five data labels of binary classification for example. In binary classification, discrete labels may be [[0, 1], [0, 1], [0, 1], [1, 0], [0, 1]],

while probabilistic labels might be [[0.2, 0.8], [0.4, 0.6], [0.5, 0.5], [0.7, 0.3], [0.1, 0.9]]. Generally, imbalance ratio (IR) [88], the ratio between the number of majority class instances and minority class instances, is used to measure the degree of class-imbalance. Accordingly, the imbalance ratio of the discrete labels [[0, 1], [0, 1], [0, 1], [1, 0], [0, 1]] can be calculated by (1+1+1+0+1) / (0+0+0+1+0), which equals to 5. Similarly, the imbalance ratio of [[0.2, 0.8], [0.4, 0.6], [0.5, 0.5], [0.7, 0.3], [0.1, 0.9]] is 31/19, calculated by (0.8+0.6+0.5+0.3+0.9) / (0.2+0.4+0.5+0.3+0.9). Obviously, 31/19 is smaller than 5, which means that for the same data, data with probabilistic labels is less imbalanced than data with discrete labels.

In view of this, probabilistic weak labels may provide a novel solution of imbalanced classification. Formally, take binary classification problem as an example. If the weak label vector of a certain data is [0.7, 0.3], it means that the data belongs to the first and second categories with probability of 0.7 and 0.3, respectively. In this way, the weak label vector of most data has a probability component in each category. Moreover, the problem of imbalanced classification will no longer exist. Thus, probabilistic labels of multiple weak supervision were proposed and tested, which can be formally represented as (2), too. But different from (2), in probabilistic labels, $0 \leq y_{ij} \leq 1$; y_{ij} C is the probability that the i-th sample belongs to the category j; For each row i, $\sum_{j=1}^{K}(y_{ij}) = 1$.

Notably, the introduce of probabilistic labels also increases the noise, which may hurt the performance of training. However, the probabilistic labels here can be generated from multiple weak supervision. That is to say, to some extent, the quality of the probabilistic labels can be guaranteed by the multiple weak supervision framework and conditional independent model, which is absolutely different from the random noise. For this, the probabilistic labels in this paper have achieved the balance of noise and imbalance implicitly by means of the proposed framework. Therefore, the probabilistic labels adopted in this paper has premise and quality assurance. As for the probabilistic labels in general sense, it does not belong to the research scope of this paper. In addition, we will carefully examine the tradeoff of imbalanced classification and noise as well as explore this problem theoretically or empirically in the future work. After all, a more general and concrete study, empirical or theoretical analysis need a new paper to represent.

Taking one step further, a bridge from multiple weak supervision to probabilistic labels is needed, which is referred to as labels integration mechanism. One natural selection is simple arithmetic mean (**SAM**). With m weak supervision sources, each sample i can generate a label vector $L_i =$

$$\left[l_{i1},\ l_{i2},\ \cdots,\ l_{im} \right] \tag{3}$$

where l_{ij} denotes label from weak supervision source j and $l_{ij} \in \{1, \cdots, k\}$; k denotes the number of classes. Based on **SAM**, probabilistic label vector Y_i can be generated:$Y_i =$

$$\left[y_{i1},\ y_{i2},\ \cdots,\ y_{ik} \right] \tag{4}$$

where each row i corresponds to one piece of data, and each column j corresponds to a category;$0 y_{ij} 1$; y_{ij} C is the probability that the i-th sample to belongs to the category j; For each row i, $\sum_{j=1}^{K}(y_{ij}) = 1$. Specifically, the arithmetic mean algorithm is shown in Algorithm 2.

Algorithm 2 Simple arithmetic mean.

Input: number of classes k; multiple weak supervision labels matrix $L_{n\times m} = \begin{bmatrix} y_{11},\ y_{12},\ \cdots,\ y_{1m} \\ y_{21},\ y_{22},\ \cdots,\ y_{22m} \\ \cdots,\ \cdots,\ y_{ij},\ \cdots \\ y_{n1},\ y_{n2},\ \cdots,\ y_{nm} \end{bmatrix}$,where n and m are the number of samples and weak supervision sources and $l_{ij} \in C = \{1, 2, ..., k\}$;

```
for i = 1 to n do
    for l = 1 to k do
        initialize y_il ← 0
        for j = 1 to m do
            if l_ij == c_l then
                y_il ← y_il + 1/m
```

Return $L_{n\times k} = \begin{bmatrix} y_{11},\ y_{12},\ \cdots,\ y_{1k} \\ y_{21},\ y_{22},\ \cdots,\ y_{22k} \\ \cdots,\ \cdots,\ y_{ij},\ \cdots \\ y_{n1},\ y_{n2},\ \cdots,\ y_{nk} \end{bmatrix}$

In fact, the multiple weak labels integration based on conditional independent model is a weighted average label integration. Based on this, the multiple weak labels integration based on conditional independent model becomes the weight determination problem of different weak supervision modes. To solve the problem of weight determination, this paper takes the "repeated calculation" correlation as an example to formally show the multiple weak labels integration based on conditional independent model. If there are m weakly supervised patterns, they are

used for unlabeled samples. When unlabeled samples meet the specific weak supervised mode, they will get weak labels, otherwise they will get *abstain* label. Therefore, in order to model the *double counting* correlation, it is necessary to ensure that the label is not *abstain*. Accordingly, this study needs to define*whether to mark* and *whether to calculate repeatedly*.

Using the above definition, the label matrix obtained by m weakly supervised modes is abbreviated as $L_{n\times m}$. For whether to mark or not,

$\phi_{i,j}^{\text{label}}(\Lambda, Y) = 1\{\Lambda_{i,j} \neq Abstain\}$

For double counting or not,

$$\phi_{i,j,k}^{correlation}(\Lambda, Y) = 1\{\Lambda_{i,j} = \Lambda_{i,k}\}, 1 \leq j \leq k \leq m$$

Accordingly, for a sample with m weakly supervised patterns, the following conditional independent model can be obtained by defining all possible *recalculation* C as $\phi_i(\Lambda, Y)$ and the corresponding weight parameter vectors $\mathrm{w} \in \mathrm{R}^{2m+|C|}$.

$$\mathrm{p}_w(\Lambda, Y) = \frac{e^{\sum_{i=1}^{n} w^T \phi_i(\Lambda, y_i)}}{Z_w}$$

where Z_w is the normalized constant. Furthermore, under the condition of only label matrix Λ and no real label vector Y, the learning of weight parameter vector has the following negative log marginal likelihood objective function.

$$\hat{w} = \arg\min_{w} -\log \sum_{Y} p_w(\Lambda, Y)$$

In this way, based on the above objective function and random gradient descent, the weight parameter vector w can be learned. Then the discrete label matrix $L_{n\times m}$ can be transformed into a more accurate probabilistic label matrix $L_{n\times k}$.

5 Methodology

As shown in Algorithm 3 and Fig. 3, the process of short text classification with multiple weak supervision mainly has five steps: (1) Knowledge Extraction, (2) Data Labeling, (3) Labels Integration, (4) Model Training and (5) Model Evaluation. It is important that the heuristic rules are domain-independent as well as the regular expressions work for any domain text classification.

Algorithm 3 Multiple weak supervision for short text classification.

Input: unlabeled short text $Train = \{x_1, x_2, \cdots, x_N\}$; small-scale hand-labeled data $Dev = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\}$, $Valid = (x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$, $Test = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$;
corpus (knowledge base) $\boldsymbol{KB} = \{(kb_1, kb_2, \cdots, kb_L), Y_{kb}\}$;
similarity threshold α; number of classes (topics) K;
number of keywords per sample M;
total number of heuristic rules and remote supervised clustering N_{lf};
minimum criteria of confusion matrix criteria *M0*

Step1 Knowledge Extraction

for $i = 1$ **to** l **do**
  for $j = 1$ **to** M **do**
    keywords $W_{ij} \leftarrow LDA\{(x_i, y_i)\}$
    label Y_{wij}, y_i
  if $y_i == Y_{kb}$ **then**
    for $k = 1$ **to** L **do**
      $\alpha_{ik} \leftarrow cos(x_i, KB_k)$
    $\alpha \leftarrow max\{\alpha_{ik}\}$

Step2 Data Labeling

for $i = 1$ **to** N **do**
  for $j = 1$ **to** L **do**
    similarity $S_i \leftarrow \max\left\{cos\left(x_i, KB_j\right)\right\}$
    if $S_i > \alpha$**then**
      $L_{ij} \leftarrow Y_{kb}$
    else
      $L_{ij} \leftarrow abstain$
  for $j = 1$ **to** M **do**
    if W_{ij} in x_i then
      $L_{ij} \leftarrow Y_{wij}$
    else
      $L_{ij} \leftarrow abstain$

Step3 Labels Integration

for $i = 1$ **to** N **do**
  for $j = 1$ **to** N_{lf} **do**
    for $k = 1$ **to** K **do**
      probabilistic label vector$L_{ik} \leftarrow Conditional\ Independent\ Model\left\{L_{ij}\right\}$

Step4 Model Training

classifier $c \leftarrow Logistic$
$\{Dense\{Train + probabilisticlabelsLp\}\}$

Step5 Model Evaluation

confusion matrix $M1 \leftarrow$ with classifier c , predict experiment in *Test*
for $i, j = 1$ **to** K **do**
  if $M1_{ij} >= M0_{ij}$ **then**
    $C \leftarrow c$
  else
    Back to **Step1**

Return $\boldsymbol{C}$

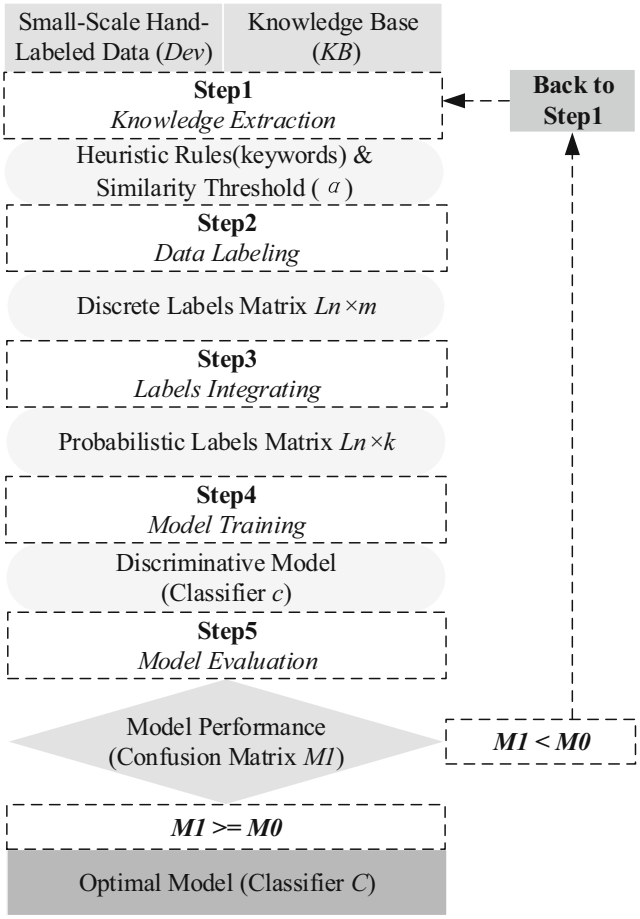

Fig. 3 Multiple Weak Supervision for Short Text Classification

(1) **Knowledge Extraction**. Knowledge extraction here refers to keywords extraction and similarity threshold calculating. However, both keywords extraction and similarity threshold determination should be based on small-scale labeled data (*Dev*), which has ground-truth label. With *Dev* and LDA (Latent Dirichlet Allocation), keywords related to specific class (topic) can be extracted effectively. Moreover, *Dev* is also the reference for the screening of distant supervision corpus. However, it is from the perspective of weak supervision and is dedicated to extracting the necessary domain knowledge to produce weak labels of data. Notably, LDA [86] was creatively applied in knowledge extraction.

(2) **Data Labeling**. Data labeling formally represents extracted knowledge and then assigns labels to unlabeled data item by item. To ensure the quality of labels, every data point can only be assigned one label if they satisfy a specific pattern. If not, the data point will only get *abstain* label. In this way, with multiple weak supervision, one data point may get more than one label. If *abstain* is also treated as a kind of label, with m weak supervision sources, one data point will get m labels. Accordingly, after n pieces of data labeling, a noisy n×m discrete labels matrix $L_{n\times m}$ will be generated. However, discrete labels matrix $L_{n\times m}$ cannot enter machine learning algorithm directly as well as cannot handle imbalanced problem, so original discrete labels matrix need to be transformed into probabilistic labels matrix.

(3) **Labels Integration**. It is assumed that the discrete label l_{ij} is generated by the true label y_i. That is, given the true label y_i, a conditional probability $P(l_{ij}|y_i)$ need to be learned. Considering that latent variable y_i cannot be observed, the label l_{ij} (other than weak supervision pattern i) is used instead. In this way, with conditional independent model, the $n\times m$ labels matrix $L_{n\times m}$ can be transformed into a n×k (k denotes the number of classes) probabilistic labels matrix Ln×k.

(4) **Model Training**. Together with *bag-of-words* (term frequency) feature vector, probabilistic labels vector can be directly used as the input of neural network for model training. In view of this, we adopt full connection layer based on *sigmoid/softmax* activation function to *train*, which can make full use of probabilistic labels.

(5) **Model Evaluation**. To evaluate the performance of classification model, test experiments were conducted on small-scaled dataset (*Test*). With the test results, we can better determine the next step. If the results meet the requirements, the output is the optimal model; Otherwise, go back to **step1** and optimize the keywords and distant supervision elements (corpus and threshold) until the model performance meets the requirements and gets the optimal model. Here, the ultimate goal of weak supervision is classification. The performance of the classification model may well illustrate the quality of weak supervision. Thus, categorical evaluation indicators such as *precision*, *recall*, and *F1-score*, rather than graph-based semi-supervised techniques, are used for model evaluation. Additionally, graph-based semi-supervised techniques were leaved for future research.

6 Experiments

6.1 Experimental settings

For simplicity and availability, the proposed method was tested to find out given topics from the title of news or tender announcements. There are two special statements here. Firstly, both oversampling and under-sampling require strong labels for large-scale training data. The proposed method is mainly based on weak labels generated by

multiple weak supervision. In other words, it is difficult to directly compare the solutions of oversampling and undersampling at the data level with the probabilistic labels resolution mechanism in this paper. Secondly, although multiple weak supervision uses both labeled and unlabeled data, it cannot be simply classified as a semi-supervised learning method. This is because multiple weak supervision can not only solve the problem of insufficient labeled data, but also solve the problems of data sparsity and imbalanced classification. Therefore, it is meaningless to compare semi-supervised learning methods such as co-training with multiple weak supervision. Given the confidentiality of the research, we will consider whether to disclose the source code and the data sets.

Datasets As we all know, public datasets, real datasets and sythetic datasets all can be used for experimental verification. For the sake of completeness and simplicity, experiments were conducted on one public dataset AG News (***AG***) [89], two sythetic datasets(synthetic binary classification dataset ***SB***, synthetic tri-classification dataset ***ST***) and one real dataset ***RD***). In particular, ***AG***'s news title and title of tender announcement were used as the sole input of model. Concretely, the basic information of ***AG***, ***SB***, ***ST*** and ***RD*** was listed in Table 3. Among them, ***SB***, ***ST*** and ***AG*** are balanced datasets, while ***RD*** is imbalanced. Moreover, all the experimental datasets used are short text with less than 50 Chinese characters or 15 English words, which indicates the data are very sparse. In addition, every data dataset includes three small-scale datasets (*Dev*, *Valid*, *Test*) with ground-truth label and large-scale unlabeled data (*Train*).

Model settings Above all, to automatically generate better weak labels, keywords matching, regular expressions and distant supervision clustering were integrated. Secondly, for simplicity and utility, the *N-gram* (feature representation) of the titles and L*ogistic Regression* (algorithm) were combined to address the challenge of data sparsity. Moreover, in order to alleviate the imbalanced classification, a fully connected neural network based on *sigmoid/softmax* activation function (Deep Logistic Regression Algorithm, *DLR*) was adopted to input probabilistic labels.

For simplicity and practicability, the *bag-of-words* of the titles is the only feature used. In addition, in order to input probabilistic labels, a fully connected neural network based on *sigmoid/softmax* activation function (Deep Logistic Regression algorithm, *DLR*) was adopted. In addition, *L2* regularization and cross-entropy loss function are used. For the sake of limited space and convenience, 3 classical algorithms (Logistic Regression (*LR*), Naïve Bayes (*NB*) and Support Vector Machine (*SVM*)) and 6 pre-training models fine tuning were tested on ***HAND*** (small-scale hand-labeled data *Dev* as training data) comparison, which will be expanded yet in the future research. After all, this article focuses more on proposing and implementing overall effective solution. To be specific, 6 pre-training models include BERT Base Chinese (BERT1) [39], BERT Base Multilingual (BERT2) [39], RoBERTa Base Chinese (RoBERTa1) [42], RoBERTa Large Chinese (RoBERTa2) [42], ERNIE Chinese (ERNIE1) [41] and ERNIE2.0 Chinese (ERNIE2) [43].

Comparison models Moreover, as an overall solution, mutiple weak supervision can solve insufficient labeled data, data sparsity and imbalanced classification. However, any one of semi-supervised learning, sampling startegy and weak supervision cannot achieve this. Moreover, accroding to No Free Lunch Theorem [91], algorithms that perform well in one domain or under certain assumptions may not necessarily be the "strongest" in another. In view of this, multiple weak supervision cannot be comapred with semi-supervised learning, sampling startegy, weak supervision and so on.

For comparison, we consider four baselines (Table 4) ***HAND*** (small-scale hand-labeled data *Dev* as training data); ***SWS*** (single type of weak supervision: only with several simple keywords matching rules); ***DET*** (with discrete

Table 3 Basic information of dataset

Dataset		*SB*	*ST*	*RD*	*AG*
Train(unlabled)	n	18000	27000	20000	60000
	AC	30.89	33.43	31.78	6.81
Dev(labled)	n	800	1200	500	100
	AC	31.78	33.91	31.77	6.81
Valid(labled)	n	600	900	500	1000
	AC	30.78	32.87	31.93	6.74
Test(labled)	n	600	900	500	1800
	AC	31.91	33.87	32.64	6.78

where n denotes the number of examples used in dataset, AC denotes the average number of characters (per sample)

Table 4 The main differences between different experiments

Experiments	Data	Weak Supervision	Labels
Baseline1: ***HAND***	*Dev*	-	PL
Baseline2: ***SWS***	*Train*	KM	PL
Baseline3: ***DET***	*Train*	KM + RE + DSC	DL
Baseline4: ***NOD***	*Train*	KM + RE	PL
Our Method: ***MWS***	*Train*	KM + RE + DSC	PL

where KM, RE, DSC are short for Keywords Matching, Regular Expressions and Distant Supervision Clustering; PL and DL are abbreviations for Probabilistic Labels and Discrete Labels

labels for training); ***NOD*** (no distant supervision sources for labeling data). ***HAND*** is used for illustrating the efficiency of large-scale data with weak labels. Compared with SWS, the strong representation ability of multiple weak supervision can be verified. ***DET*** can highlight the role of probabilistic labels in imbalanced classification problem, while ***NOD*** can validate the importance of distant supervision clustering.

6.2 Experimental results

It should be noted that the synthetic dataset ***SB*** and ***ST*** were strictly selected by keyword matching. Therefore, the heuristic rules of simple keyword matching are consistent with ***SB*** and ***ST***, and the experimental results in ***SB*** and ***ST*** may well be similar to multiple weak supervision method. Notably, the bold emphasis in Tables 5, 6, 7, 8 and 9 are used to highlight the best experimental results.

(1)*HAND* comparison From Table 5, the results of synthetic dataset ***SB*** and ***ST*** on *Dev* and *Train* were similar, both get above 95% score. This is because the synthetic datasets ***SB*** and ***ST*** were strictly selected based on keyword matching pattern. But it also suggests that the process of model training is translating weak supervision strategies into machine learning models, or integrating several weak classifiers into one strong classifier, intellectually similar to stacking [90]. Notably, the results of datasets ***RD*** and ***AG*** well illustrate the huge advantages of expanding training samples with multiple weak supervision and improving training effect. Particularly, in ***RD***, *F1-score* was improved by an average of 32 percentage points.

In addition, considering the relative poor perfrmance of 3 classical algorithms on dataset ***RD***, *fine tuning* experiments were also added. To be specific, 6 pre-training models were adopted, which include BERT Base Chinese (BERT1) [39], BERT Base Multilingual (BERT2) [39], RoBERTa Base Chinese (Ro-BERTa1) [42], RoBERTa Large Chinese (RoBERTa2) [42], ERNIE Chinese (ERNIE1) [41] and ERNIE2.0 Chinese (ERNIE2) [43]. To be specific, the experimental results of *fine tuning* are shown in Table 6. According to Table 6, the *recall* and *F1-Score* of ***MWS*** are better than the fine tuning results of all the six pre-training models. In terms of *precision*, ***MWS*** is also no less than four pre-training models. This is not contrary to the effectiveness of *fine tuning* on small-scale strongly labeled data sets. This

Table 5 Results between *Dev* and *Train*

Training dataset		*Dev*			*Train*
Algorithm		*LR*	*NB*	*SVM*	*DLR*
SB	*Precision*	0.98	0.95	0.98	0.98
	Recall	0.98	0.95	0.98	0.98
	F1-score	0.98	0.95	0.98	0.98
ST	*Precision*	0.99	0.95	0.99	0.98
	Recall	0.99	0.95	0.99	0.98
	F1-score	0.99	0.94	0.99	0.98
RD	*Precision*	0.51	0.50	0.60	**0.86**
	Recall	0.71	0.46	0.74	**0.86**
	F1-score	0.51	0.48	0.63	**0.86**
AG	*Precision*	0.66	0.70	0.69	**0.75**
	Recall	0.67	0.70	0.70	**0.80**
	F1-score	0.65	0.70	0.69	**0.77**

Table 6 Results between ***MWS*** and Pre-Training Model ***Fine Tuning*** on ***RD***

Methods	Models	*Precision*	*Recall*	*F1-score*
Fine Tuning	*BERT1*	0.77	0.57	0.61
	BERT2	0.76	0.54	0.55
	RoBERTa1	**0.86**	0.63	0.68
	RoBERTa2	0.79	0.59	0.62
	ERNIE1	0.82	0.56	0.59
	ERNIE2	0.84	0.54	0.55
MWS	*DLR*	0.82	**0.74**	**0.78**

is because the small-scale strong labeled data set used for fine tuning becomes the large-scale weak labeled data set. After all, *fine tuning* relies on strong labeled data, which cannot be provided by weak supervision.

(2)*SWS* comparison Table 6 shows that, with single type of weak supervision (***SWS***), the performance of ***SB*** and ***ST*** is so good that there is little room for improvement. Therefore, multiple weak supervision (***MWS***) was only tested in ***RD*** and ***AG***. From Table 7, the performance of ***MWS*** is significantly better than that of ***SWS***. This fully illustrates the obvious advantages of ***MWS*** over ***SWS***, and proves the effectiveness of ***MWS*** method. In particular, with the help of ***MWS***, the *F1-score* in ***RD*** has increased by 2%.

(3)*DET* comparison ***RD*** covers a wide variety of topics, but we only try to find the topic we care about. In view of this, it is a binary classification problem. Moreover, compared with uninterested topics, the proportion of topics we care about are very low. That is, ***RD*** is imbalanced, while ***SB*** and ***ST*** are balanced. In order to verify the effect of probabilistic labels on solving imbalanced classification problem, we carried out the control test on imbalanced dataset **RD** based on probabilistic labels and discrete labels respectively. The results are shown in Table 8.

The results on imbalanced dataset ***RD*** (Table 8) fully illustrate the advantages of probabilistic labels in solving imbalanced classification problems compared with discrete labels. Specifically, probabilistic labels provide a 9% improvement of *F1-score* on *Test*. Table 8 shows that the probabilistic labels can improve the classification performance of minority class remarkably. Compared to 2% of majority class, the *F1-score* of minority class was improved by 16% with the help of probabilistic labels. In a sense, probabilistic labels, or multiple weak supervision, might provide a new possibility for solving imbalanced classification problem.

Table 7 Results between ***SWS*** and ***MWS***

	RD		*AG*	
	SWS	*MWS*	*SWS*	*MWS*
Precision	0.78	**0.86**	0.75	**0.75**
Recall	0.77	**0.86**	0.77	**0.80**
F1-score	0.78	**0.86**	0.75	**0.77**

where ***SWS*** and ***MWS*** are abbreviations for single type of weak supervision and multiple weak supervision, respectively

(4) *NOD* comparison Experimental results show that, with weak labels form heuristic rules, the performance of ***SB*** and ***ST*** is good enough for application. Therefore, distant supervision clustering was only tested on datasets ***RD*** and ***AG***. In detail, the experimental results are listed on Table 9.

In ***WD***, the *recall* score of ***RD*** was improved by 4% without reduction in *precision* score. This suggests that similarity threshold can act as the regulator of *recall*. Therefore, adjusting similarity threshold can meet different application needs , which is of great significance in academia and industry.

To sum up, we have the following observations.

(1) While multiple weak supervision expands the labeled dataset, it also alleviates data sparsity of short text, thus improving the performance of the classifier.
(2) With conditional independent model, weak labels provided by multiple weak supervision have higher accuracy and coverage than those provided by single type of weak supervision.

Table 8 ***DET*** experimental results on **RD**

Performance		Minority	Majority	Macro	Weighted
DL	*Precision*	0.70	0.94	0.82	0.92
	Recall	0.49	0.98	0.73	0.92
	F1-score	0.58	0.96	0.77	0.92
PL	*Precision*	**0.74**	0.98	**0.86**	0.96
	Recall	**0.74**	0.98	**0.86**	0.96
	F1-score	**0.74**	0.98	**0.86**	0.96

DL and ***PL*** are abbreviations for discrete labels and probabilistic labels, respectively

Table 9 Experimental results between ***NOD*** and ***WD***

Dataset	RD		AG	
Weak Supervision	*NOD*	*WD*	*NOD*	*WD*
Precision	0.86	0.86	**0.76**	0.75
Recall	0.86	**0.90**	0.76	**0.80**
F1-score	0.86	**0.88**	0.76	**0.77**

NOD and ***WD*** are abbreviations for no distant supervision clustering and with distant supervision clustering, respectively

(3) Probabilistic labels may provide a new solution for imbalanced classification problem. Notably, probabilistic labels should base on reliable multiple weak supervision.

The similarity threshold can be the regulator of *recall*. That is, distant supervision clustering can be used to represent tacit knowledge and improve *recall* score.

(4) For multiple weak supervision, LDA can be used to extract explicit knowledge (keywords) of heuristic rules efficiently.

Additionally, based on the comparison results of the above four experiments, the effectiveness of the proposed framework in solving labels shortage, data sparsity and imbalanced classification wholly has also been fully illustrated. In general, the proposed framework can be used for short text classification of any domain. Notably, the main differences among different domains are the keywords pattern, external corpus and similarity threshold. That is, with proper keywords and relevant corpus, there is little difference in the performance of the proposed framework in different areas of the short text classification.

7 Conclusion

To address the labels bottleneck, data sparsity and imbalanced classification in short text classification simultaneously, multiple weak supervision was designed. With multiple weak supervision, implicit knowledge and tacit knowledge can be used to generate weak labels automatically. Furthermore, based on weak labels and conditional independent model, probabilistic labels and effective imbalanced classification model can be trained. What makes it reasonable is that implicit knowledge and tacit knowledge can provide enough diversity for labels integration. Specifically, our work has the following four contributions:

(1) Multiple Weak Supervision Sources: Multiple weak supervision sources, covering explicit knowledge and tacit knowledge, were creatively introduced to label training data. Taking short text classification as an example, multiple weak supervision sources can be *simple keywords matching*, *regular expressions* and *distant supervision clustering*.

(2) Probabilistic Labels for Imbalanced Classification: Experimental results show that, the probabilistic labels generated by conditional independent model can effectively solve the imbalanced text classification problem. This may provide a new solution to imbalanced classification, which has troubled industry workers and researchers for years.

(3) Combining Distant Supervision with Clustering: Different from common alignment strategy, distant supervision was combined with clustering for generating weak labels and improving the coverage. In this way, distant supervision clustering was proposed, which can make full use of small-scale hand-labeled data and does not need explicit knowledge extraction. With distant supervision clustering, tacit knowledge, which is hard to represent, can be included in knowledge base (corpus) and similarity threshold easily.

Notably, the similarity threshold of distant supervision clustering can be used as the regulator of *recall*. In practical applications, this is of great significance for applying weak supervision to meet different needs of *recall* score. That is, if the clustering corpus and similarity threshold can be selected well, the *recall* and *F1-score* could be improved with little effect on precision.

(4) LDA for Knowledge Extraction: Latent Dirichlet Allocation (LDA) was introduced to extract keywords of specific topic, which is the foundation of weak supervision. Moreover, LDA can effectively prevent over-fitting, which is also very simple and useful.

Despite of this, there are still many limitations in this paper. In future, we will further study the knowledge extraction methods (such as LDA), expand weak supervision sources and seek more theoretical analysis to validate the multiple weak supervision method.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Ratner A et al (2017) Snorkel: Rapid Training Data Creation with Weak Supervision. Proc VLDB Endowment 11(3):269–282
2. Sun C et al (2017) Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In: 2017 IEEE International Conference on Computer Vision (ICCV)
3. Bach SH et al (2019) Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale. Proc ACM SIGMOD Int Conf Manag Data 2019:362–375
4. Zhou Z (2018) A brief introduction to weakly supervised learning. Ntl Sci Rev 5(1):44–53
5. Ratner A et al (2016) Data Programming: Creating Large Training Sets, Quickly. Adv Neural Inf Process Syst 29:3567–3575
6. Zhu X, Goldberg AB (2009) Introduction to Semi-Supervised Learning. Synthesis Lect Artif Intell Mach Learn 3(1):130
7. Alfonseca E et al (2012) Pattern learning for relation extraction with a hierarchical topic model. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2. Association for Computational Linguistics, Jeju Island, pp 54–59
8. Augenstein I, Maynard D, Ciravegna F (2014) Relation Extraction from the Web Using Distant Supervision. In: International Conference on Knowledge Engineering and Knowledge Management
9. Mintz M et al (2009) Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. Association for Computational Linguistics, Suntec, pp 1003–1011
10. Quinn AJ, Bederson BB (2011) Human computation: a survey and taxonomy of a growing field. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, Vancouver, pp 1403–1412
11. Yuen M, King I, Leung K (2011) A Survey of Crowdsourcing Systems. In: 2011 IEEE Third International Conference on Privacy: Security Risk and Trust and 2011. IEEE Third International Conference on Social Computing
12. Rekatsinas T et al (2017) HoloClean: holistic data repairs with probabilistic inference. Proc VLDB Endow 10(11):1190–1201
13. Sa CD et al (2016) DeepDive: Declarative Knowledge Base Construction. SIGMOD Rec 45(1):60–67
14. Liang P, Jordan MI, Klein D (2009) Learning from measurements in exponential families. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, Montreal, pp 641–648
15. Mann GS, McCallum A (2010) Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. J Mach Learn Res 11:955–984
16. Stewart R, Ermon S (2016) Label-Free Supervision of Neural Networks with Physics and Domain Knowledge. Thirty-First Aaai Conference on Artificial Intelligence, pp 7
17. Zaidan OF, Eisner J (2008) Modeling annotators: a generative approach to learning from annotator rationales. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Honolulu, pp 31–40
18. Yao W, Liu J, Cai Z (2018) Personal Attributes Extraction in Chinese Text Based on Distant-Supervision and LSTM. In: Advances in Computer Science and Ubiquitous Computing. Springer Singapore, Singapore
19. Shi Y, Xiao Y, Niu L (2019) A Brief Survey of Relation Extraction Based on Distant Supervision in Computational Science – ICCS 2019. Springer International Publishing, Cham
20. Batista-Navarro R, Hawkins O (2019) Topic Modelling vs Distant Supervision: A Comparative Evaluation Based on the Classification of Parliamentary Enquiries. In: Digital Libraries for Open Knowledge. Springer International Publishing, Cham
21. Krause S et al (2012) Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web in The Semantic Web – ISWC 2012. Springer, Berlin
22. Heist N, Paulheim H (2017) Language-Agnostic Relation Extraction from Wikipedia Abstracts in The Semantic Web – ISWC 2017. Springer International Publishing, Cham
23. Auer S et al (2007) DBpedia: A Nucleus for a Web of Open Data. In: The Semantic Web. Springer, Berlin
24. Doan A, Ramakrishnan R, Halevy AY (2011) Crowdsourcing systems on the World-Wide Web. Commun ACM 54(4):86–96
25. Haralabopoulos G et al (2019) Paid Crowdsourcing, Low Income Contributors, and Subjectivity. In: Artificial Intelligence Applications and Innovations. Springer International Publishing, Cham
26. Nowak S et al (2010) How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on Multimedia information retrieval. ACM, Philadelphia, pp 557–566
27. Redi JA et al (2013) Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In: Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia. ACM, Barcelona, pp 29–34
28. Vondrick C, Patterson D, Ramanan D (2013) Efficiently Scaling up Crowdsourced Video Annotation. Int J Comput Vis 101(1):184–204
29. Alonso O, Rose DE, Stewart B (2008) Crowdsourcing for relevance evaluation. SIGIR Forum 42(2):9–15
30. Akkaya C et al (2010) Amazon Mechanical Turk for subjectivity word sense disambiguation. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, Los Angeles, pp 195–203
31. Callison-Burch C, Dredze M (2010) Creating speech and language data with Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, Los Angeles, pp 1–12
32. Gao Q, Vogel S (2010) Consensus versus expertise: a case study of word alignment with Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, Los Angeles, pp 30–34
33. Nassar L, Karray F (2019) Overview of the crowdsourcing process. Knowl Inf Syst 60(1):1–24
34. Bach SH et al (2017) Learning the Structure of Generative Models without Labeled Data. Proc Mach Learn Res 70:273–82
35. Wang H et al (2019) An Empirical Study of Heuristic Rules on the Performance of Satellite TT&C Scheduling Algorithms. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)
36. Peters M et al (2018) Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics
37. Radford A et al (2018) Improving language understanding by generative pre-training
38. Devlin J et al (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding in NAACL-HLT

39. Yang ZL et al (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Wallach H et al (eds) Advances in Neural Information Processing Systems
40. Diao S et al (2019) ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. arXiv e-prints
41. Sun Y et al (2019) ERNIE: Enhanced Representation through Knowledge Integration. arXiv:1904.09223
42. Liu Y et al (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach
43. Sun Y, Sun Y et al (2020) ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence
44. da Silva NFF, Hruschka ER, Hruschka ER (2014) Tweet sentiment analysis with classifier ensembles. Decis Support Syst 66:170–179
45. Heap B et al (2017) Word Vector Enrichment of Low Frequency Words in the Bag-of-Words Model for Short Text Multi-class Classification Problems
46. Alsmadi I, Gan KH (2019) Review of short-text classification. Int J Web Inf Syst 15(2):155–182
47. Diao S et al (2019) ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. arXiv e-prints
48. Allahyari M et al (2017) A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques
49. Zhuo Z et al (2020) News Text Topic Clustering Optimized Method Based on TF-IDF Algorithm on Spark. Comput Mater Cont 62(1):217–231
50. Kadhim AI (2019) Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF. In: 2019 International Conference on Advanced Science and Engineering (ICOASE)
51. Deng X et al (2019) Feature selection for text classification: A review. Multimed Tools Appl 78(3):3797–3816
52. Ge S et al (2014) Short Text Classification: A Survey. J Multimed 9(5):635–643
53. Ostrowski DA (2014) Feature Selection for Twitter Classification in 2014. IEEE International Conference on Semantic Computing
54. El Akadi A et al (2011) A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowl Inf Syst 26(3):487–500
55. Meng J, Lin H, Yu Y (2011) A two-stage feature selection method for text categorization. Comput Math Appl 62(7):2793–2800
56. Mundra PA, Rajapakse JC (2010) SVM-RFE With MRMR Filter for Gene Selection. IEEE Trans NanoBiosci 9(1):31–37
57. Tenenhaus M et al (2005) PLS path modeling. Comput Stat Data Anal 48(1):159–205
58. Deerwester S et al (1990) Indexing by latent semantic analysis 41(6):391–407
59. Zareapoor M, Seeja K. J. I. J. o. I. E., Business E (2015) Feature extraction or feature selection for text classification: A case study on phishing email detection 7(2):60
60. Bharti KK, Singh PK (2015) Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. Expert Syst Appl 42(6):3105–3114
61. Chen J et al (2009) Feature selection for text classification with Naïve Bayes. Expert Syst Appl 36(3, Part 1):5432–5435
62. Wang M, Lin L, Wang F (2013) Improving Short Text Classification through Better Feature Space Selection in 2013. Ninth International Conference on Computational Intelligence and Security
63. Weissbock J, Esmin AA, Inkpen D (2013) Using external information for classifying tweets. In: 2013 Brazilian Conference on Intelligent Systems. IEEE
64. Goyal S, Parveen S (2015) Improved feature selection for better classification in twitter. Int J Comput Appl 122(1)
65. Rosa H, Batista F, Carvalho JP (2014) Twitter Topic Fuzzy Fingerprints in 2014. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
66. Yin C et al (2015) A New SVM Method for Short Text Classification Based on Semi-Supervised Learning. In: 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS)
67. Kotsianti SB, Kanellopoulos D (2007) Combining Bagging, Boosting and Dagging for Classification Problems. Springer, Berlin
68. Rogati M, Yang Y (2002) High-performing feature selection for text classification, Inproceedings of the eleventh international conference on Information and knowledge management. Association for Computing Machinery, McLean, pp 659–661
69. Forman G (2003) An Extensive Empirical Study of Feature Selection Metrics for Text Classification. J Mach Learn Res 3(7/8):1289–1305
70. Chawla N, Japkowicz N, Kolcz A (2003) Workshop learning from imbalanced data sets II. In: Proceedings of Int'l Conf Machine Learning
71. Chawla N, Japkowicz N, Kotcz A (2004) Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor Newsl 6(1):1–6
72. Wang S et al (2017) Proceedings of the IJCAI 2017 Workshop on Learning in the Presence of imbalanced classification and Concept Drift (LPCICD'17) arXiv e-prints
73. Chawla N et al (2002) SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res (JAIR) 16:321–357
74. Han H, Wang W-Y, Mao B-H (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Springer, Berlin
75. Haibo H et al (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)
76. Feng Y, Zhou M (2020) X Tong Imbalanced classification: an objective-oriented review. arXiv e-prints
77. Liu X, Wu J, Zhou Z (2009) Exploratory Undersampling for Class-Imbalance Learning. IEEE Trans Syst Man Cybern Part B (Cybern) 39(2):539–550
78. Luqyana WA, Ahmadie BL, Supianto AA (2019) K-Nearest Neighbors Undersampling as Balancing Data for Cyber Troll Detection. In: 2019 International Conference on Sustainable Information Engineering and Technology (SIET)
79. López V. et al (2013) An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141
80. Liang G (2013) An Effective Method for Imbalanced Time Series Classification: Hybrid Sampling
81. Gan D et al (2020) Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Comput Ind Eng 140:106266
82. Błaszczyński J, Stefanowski J (2015) Neighbourhood sampling in bagging for imbalanced data. Neurocomputing 150:529–542
83. Yuan Z, Zhao P (2019) An Improved Ensemble Learning for Imbalanced Data Classification. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)
84. Padurariu C, Breaban ME (2019) Dealing with Data Imbalance in Text Classification. Procedia Comput Sci 159:736–745
85. Nonaka I (1994) A Dynamic Theory of Organizational Knowledge Creation. Organ Sci 5(1):14–37
86. Blei DM et al (2003) Latent Dirichlet Allocation. J Mach Learn Res 3:993–1022

87. Zhang Y et al (2020) Keywords extraction with deep neural network model. Neurocomputing 383:113–121
88. Orriols-Puig A, Bernadó-Mansilla E (2009) Evolutionary rule-based systems for imbalanced data sets. Soft Comput 13:213–225
89. Corso GMD, Gullí A, Romani F (2005) Ranking a stream of news, Inproceedings of the 14th international conference on World Wide Web. Association for Computing Machinery, Chiba, pp 97–106
90. Wolpert D (1992) Stacked Generalization. Neural Netw 5:241–259
91. Wolpert D, Macready W (1997) The No Free Lunch Theorems for Optimization. IEEE Trans Evol Comput 1:67–82

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.