J Inf Process Syst, Vol.17, No.6, pp.1057~1070, December 2021
https://doi.org/10.3745/JIPS.01.0081

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Case-Related News Filtering via Topic-Enhanced Positive-Unlabeled Learning

Guanwen Wang*, Zhengtao Yu**, Yantuan Xian*, and Yu Zhang*

Abstract

Case-related news filtering is crucial in legal text mining and divides news into case-related and case-unrelated categories. Because case-related news originates from various fields and has different writing styles, it is difficult to establish complete filtering rules or keywords for data collection. In addition, the labeled corpus for case-related news is sparse; therefore, to train a high-performance classification model, it is necessary to annotate the corpus. To address this challenge, we propose topic-enhanced positive-unlabeled learning, which selects positive and negative samples guided by topics. Specifically, a topic model based on a variational autoencoder (VAE) is trained to extract topics from unlabeled samples. By using these topics in the iterative process of positive-unlabeled (PU) learning, the accuracy of identifying case-related news can be improved. From the experimental results, it can be observed that the F1 value of our method on the test set is 1.8% higher than that of the PU learning baseline model. In addition, our method is more robust with low initial samples and high iterations, and compared with advanced PU learning baselines such as nnPU and I-PU, we obtain a 1.1% higher F1 value, which indicates that our method can effectively identify case-related news.

Keywords

Case-Related News, Iterative Training, Positive-Unlabeled Learning, Topic

1. Introduction

Case-related news refers to news reports on cases or potential cases. Case-related news filtering aims to quickly extract the required news related to the case from the news data, which is significant for public opinion supervision, prevention, and control.

There are two classic methods of filtering news [1,2]: keyword-based and machine-learning filtering. Initially, researchers matched the news text by collecting domain-related keywords such as Knuth-Morris-Pratt (KMP) [3] and Sunday [4]. However, keyword-based filtering relies heavily on keyword dictionary completeness and may result in high-precision and low-recall situations with a low degree of completeness. Owing to complicated and changeable case-related news reports, it is challenging to construct a complete keyword dictionary. The machine-learning method is an effective solution for filtering news that makes assumptions about the data distribution of the news categories involved, such as support vector machine (SVM) [5] and decision trees [6]. However, these methods depend on

Manuscript received January 22, 2021; first revision March 25, 2021; accepted May 30, 2021.

Corresponding Author: Yantuan Xian (xianyt@kust.edu.cn)

* College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China (2026045953@qq.com, xianyt@kust.edu.cn, 2237276545@qq.com)

** Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, China (ztyu@hotmail.com)

handcrafted feature functions and often satisfy the curse of dimensionality. The deep neural network method learns vector representation from the text without handcrafted feature functions, to alleviate dimensional disasters caused by statistical methods; however, this requires a large amount of labeled data. The lack of labeled training data will challenge text-filtering methods to achieve desirable results. In real-world applications of case-related news filters, there are only a negligible number of labeled case-related news samples; however, it is easy to collect a large amount of unlabeled news. Therefore, we focus on finding a way to improve filtering performance by mining unlabeled news samples.

Machine learning using only positive samples is a common challenge in the field of fake reviews and recommendations. To address this challenge, researchers have proposed a positive-unlabeled (PU) learning algorithm based on positive and unlabeled samples [7]. The standard PU learning method is a two-step process [8]. First, it identifies reliable negative samples and trains a classifier using these negative and the original positive samples. Next, the trained classifier is utilized to score unlabeled examples and select reliable positive and negative samples as new training samples. By repeating the above steps, we can obtain a classifier with a higher accuracy. This provides an effective training method for datasets with only positive samples. However, the final performance of the classifier largely depends on the initial reliable labeled data. Regarding a negligible number of initially labeled data samples, the iteratively trained accuracy of the classifier is insufficient. The positive and negative samples obtained may not be reliable, and there may be more misclassified data. These unreliable data will be added later in the iterative training process, resulting in increasingly misclassified data, which will seriously affect the performance of the final classifier. In this study, this phenomenon is referred to as "error accumulation." Error accumulation is caused by insufficient use of text information in the PU learning process. Case-related news contains hidden information, such as topic information. Topic information is a unique news attribute that can be obtained using unsupervised methods [9]. By leveraging this information in the initial training and subsequent iterations of PU learning, "error accumulation" can be effectively alleviated.

In this study, we propose a PU learning method combined with topic information to filter case-related news, which utilizes a negligible number of labeled case-related news samples. Our method extracts topic information from the labeled and unlabeled case-related news datasets through the unsupervised pre-training topic model and adds the topic information to the initial training and subsequent iterative training processes of PU learning. With our method, more case-related topic information can be used when the initial labeled samples are small, and topic enhancements are performed in the subsequent iteration of the training process. This allows the classifier to be trained in each iteration to obtain reliable positive and negative sample data from unlabeled data and improve the performance of the final case-related news classifier.

To conclude, the main contributions are as follows:

- We applied the PU learning method to the case-related news filtering task, which effectively tackles filtering case-related news under a negligible number of manual annotations.
- We extracted the topic information using the variational autoencoder (VAE) topic model and enhanced the text representation by learned topic representation within the PU learning training process, which significantly stabilized the negative sample selection.
- We constructed a dataset of case-related news and used our method to conduct the experiments. This indicates that our method achieves better results than the PU learning method without topic enhancement.

2. Related Work

Recently, in the recommendation system and spam review filtering domain, researchers have made a series of achievements in the classification of only positive samples, which can be summarized into the following three methods [10].

The one-class classification approach uses only positive sample data in the training set. Its core idea is to construct a minimum region that approximately covers the training set, whereas instances outside the region belong to negative samples. Manevitz and Yousef [11] proposed a one-class SVM classification method for text classification. Because this method completely ignores the unlabeled dataset, hidden classification information in the unlabeled dataset is lost. When there are reliable negative samples in the unlabeled dataset, the model is prone to overfitting because it ignores this valuable information.

The two-step approach uses positive samples and samples from an unlabeled dataset to build the final classifier. It mainly comprises two steps. First, it uses a heuristic strategy to identify negative sample data with high credibility in unlabeled data. Second, these negative sample data are combined with existing positive samples to form new training samples, and existing classification methods are used to study classifiers in the new training samples. The above algorithm framework can use iterative training to train the classifiers. The disadvantage of this method is that the performance of the final classifier depends significantly on the initial reliable sample data. If the scale is negligible or the sample quality is not high, the performance of the classifier is limited.

In addition, some researchers have considered using positive samples and all unlabeled samples for training. The core idea of these methods is to establish a binary classifier to determine the labels of the unlabeled samples, convert the unlabeled sample dataset into labeled data, and train with known positive samples. Ren et al. [12] proposed a PU-based learning algorithm applied to fake reviews. Li et al. [13] applied the conventional PU challenge to a streaming data environment and proposed a PU learning algorithm based on clustering. Xiao et al. [14] proposed a PU learning algorithm based on the similarity. First, positive samples were utilized to extract reliable negative samples from the unlabeled sample dataset. Further, based on the positive and extracted negative samples, the probability that the remaining unlabeled samples belong to positive and counterexamples was calculated, and an SVM classifier with probability weight was established based on the above data.

The topic model mainly adopts Gibb's sampling, variational inference, non-negative matrix factorization, and other machine-learning algorithms to infer potential topics from high-dimensional sparse text feature spaces [15]. VAE is an encoding-decoding network proposed by Kingma and Welling [16] in 2014. Regarding topic modeling, Miao et al. [17] first attempted to use VAE to build a neural variational document model, and on this basis considered topic-word distribution, thus forming a topic model based on a neural self-encoder structure [18–20]. The VAE is an unsupervised model; that is, it does not need to label the data but only constructs the optimization function and trains the model by reconstructing the data, which is very suitable for the application scenario of this task. Therefore, we choose VAE to model the topic of case-related news and integrate the extracted unsupervised topic into the iterative process of PU learning to improve the "error accumulation" challenge existing in PU learning and improve the performance of case-related news filtering. Therefore, we choose VAE to model the topic of case-related news and integrate the extracted unsupervised topic into the iterative process of PU learning, to improve the "error accumulation" challenge existing in PU learning and the performance of case-related news filtering.

3. Case-Related News Filtering Method

Based on PU learning under the framework of a neural network, we propose a PU learning method with topic enhancement, and combine it with a case-related news filtering model to improve the performance of case-related news filtering. This method can be divided into training, prediction, and iterative processes. The specific method is illustrated in Fig. 1.

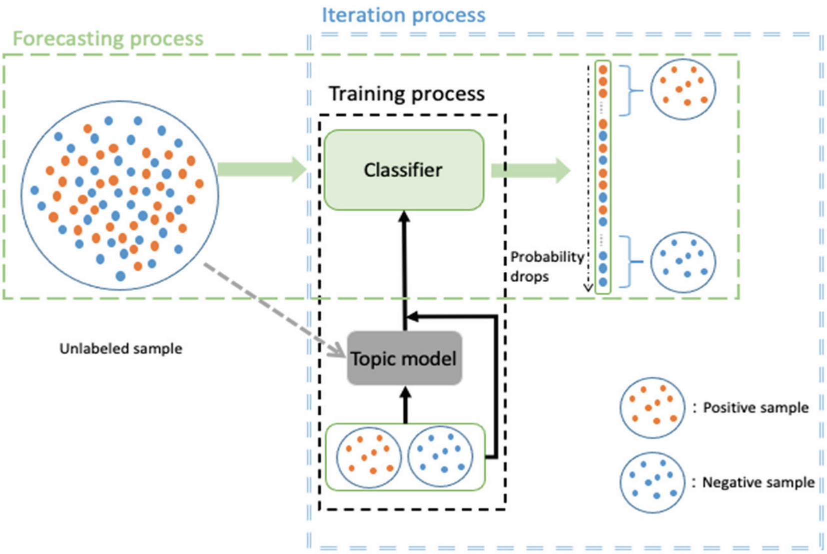


Fig. 1. A classification model of case-related news that integrates topic to enhance PU learning.

3.1 VAE Topic Model

The VAE is an unsupervised document generation model, as shown in Fig.2. Its purpose is to extract potential features from the word vector space of documents, which we refer to as topic features. Gururangan et al. [9] used VAE to extract topics to assist text classification tasks. Referring to previous work and VAE principles, we implemented this VAE structure and used the entire case-related news dataset for unsupervised pre-training.

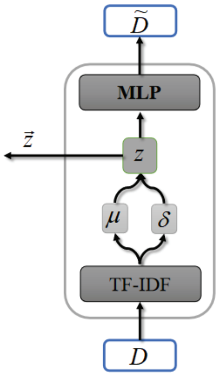


Fig. 2. VAE topic model.

The VAE architecture is an encoder-decoder architecture. In the encoder, the input is compressed into a potential distribution Z, whereas the decoder reconstructs the input signal D by sampling according to the distribution of Z in the data potential space.

$$P(D) = \sum_{Z} P(D|Z)P(Z) \tag{1}$$

$$\log P\left(Z\middle|d^{(i)}\right) = \log N(z; u^{(i)}, \delta^{2(i)}I) \tag{2}$$

where $d^{(i)}$ represents a real sample in D, and μ and δ^2 are generated by $d^{(i)}$ through a neural network. From the obtained $\mu^{(i)}$ and $\delta^{2(i)}$, the distribution $P(Z^{(i)}|d^{(i)})$ corresponding to each $d^{(i)}$ can be obtained, and $\tilde{d}^{(i)}$ can be reconstructed by decoding network $\tilde{d}^{(i)} = Decode(Z^{(i)})$. We used multilayer perception (MLP) for the generation of μ and δ^2 and the implementation of decoding network decoding.

$$Z^{(i)} = \text{Encode}\left(d^{(i)}\right), i \in R^m \tag{3}$$

$$\tilde{d}^{(i)} = \text{Decode}(Z^{(i)}), i \in R^m \tag{4}$$

where m represents the default number of potential topics. After the calculation above, the potential topic distribution of in this study can be expressed as $\vec{z} = \{\mathrm{Z}^{(1)}, \mathrm{Z}^{(2)}, \dots, \mathrm{Z}^{(\mathrm{m})}\}$.

$$L = \sum_{i=1}^{s} -\mathrm{E}_{Z-p(Z|d^{(i)})}\left[\log p\left(d^{(i)}\middle|Z\right)\right] + KL(p(Z|d^{(i)})||N(0,I)) \tag{5}$$

3.2 Topic-Enhanced Positive-Unlabeled Learning

The PU learning process comprises three steps: training, prediction, and iterative processes. In the training and iterative processes, we guide and enhance the topics obtained by the unsupervised topic model. Because we have mainly improved the training process, we will introduce it.

3.2.1 Training and predicting of PU learning

The training classifier is the main training process of the PU learning method, and an unsupervised topic model is used for enhancement. There are only a small amount of case-related news data and unmarked data in the dataset. Therefore, extracting reliable unrelated news data from unlabeled data is the first problem to be solved by the algorithm, and the initial training is executed in combination with the existing data.

To extract reliable non-case-related news data from unlabeled data, we use an improved version of the I-DNF [21] algorithm. First, the non-case-related news set should be extracted using the different frequencies of text features in the case-related news and unlabeled sample sets. We used I-DNF to obtain counterexamples of the same scale as the initial case-related news, and further trained the initial classifier. Regarding the construction of classifiers, a variety of machine-learning algorithms or deep networks can be used. Our method uses embedding and the network structure of bidirectional long-term and short-term memory network (LSTM) as classifiers [22].

First, the embedding network layer embeds the sparse representation of the raw data into a high-dimensional space and then forms a dense matrix with semantic representation. The continuous bag-of-words (CBOW) network structure of word2vec is used to build the embedding network layer [23]. First, the text is segmented, and the position code of each word is obtained according to the dictionary. The word-embedding vector $\vec{x}$ of each word is obtained through embedding and combined to obtain the word vector of the text $X = \{\vec{x_1}, \vec{x_2}, \dots, \vec{x_n}\} \in R^{n*v}$, where n represents the length of the news text, and v is the word vector dimension. In addition, the input text is passed through the VAE theme model to obtain the theme vector $\vec{z} \in R^{m}$ of the news text, where m is the preset number of themes. After obtaining the two types of encoded information, the news topic vector $\vec{z}$ is used to guide the word-embedding vector X. Because the topic vector obtained by the case-related topic model is a vector with a shape of $1^{*}m$, we made n copies of it and spliced them into the word-embedding vector X respectively, and the new matrix X' formed is the news decode vector integrated with the topic vector.

$$X' = \{\vec{x_1} + \vec{z}, \vec{x_2} + \vec{z}, \dots, \vec{x_n} + \vec{z},\} \in R^{n*(v+m)} \tag{6}$$

Bidirectional long short-term memory (BiLSTM) is a temporal network that can model data well onto close sequential relationships, such as text. It has three gating mechanisms—the forget gate, input gate, and output gate—to alleviate the vanishing gradient and capture the long-term dependency on contexts. We add the news decode vector with the topic and model its context through the BiLSTM network layer to obtain the news semantic representation vector. The specific formula is as follows:

$$H = \mathrm{BiLSTM}(X') \in R^{q*n*(v+m)} \tag{7}$$

$$`y = \mathrm{softmax}(\mathrm{MLP}(H)) \tag{8}$$

where H is the sentence vector encoded by the BiLSTM, q is the hidden layer dimension of the BiLSTM, and y represents the final probability output. Our method predicts the remaining unlabeled data using a classifier model. In addition, the probability of the prediction results for unlabeled news is sorted from high to low. In each prediction, the data with the highest probability will be obtained according to a certain iteration step as reliable case-related news samples, and the data with the lower probability as reliable negative samples, which will be removed from the unlabeled samples and added to the training data for the subsequent iterative training process. When case-related news samples are predicted, the subsequent samples are all negative.

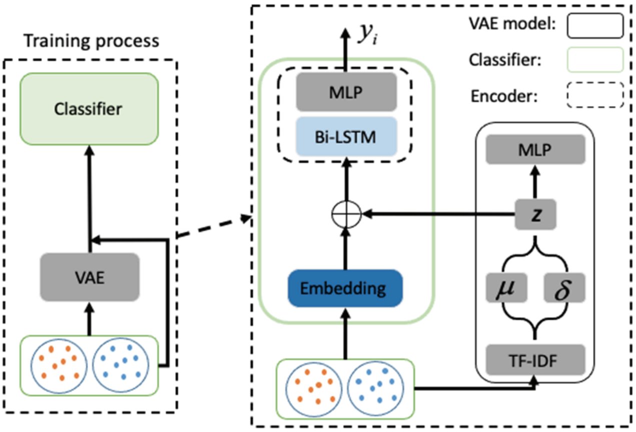


Fig. 3. The training process of PU learning.

3.2.2 PU learning iterative algorithm

Fig. 3 illustrates the iteration process of PU learning, where Steps 1–3 describe the pre-training of the topic model and the process of obtaining positive and negative samples for the first time; Steps 4–16 describe the iteration process of the entire PU learning. PU learning retrains the classifier on the newly obtained training set and repeats the entire prediction and training process after completing the initial training and prediction processes. There is no difference between the training, iterative process prediction, and the initial training prediction. The difference is that the number of unlabeled data points decreases, and the training set increases after each iteration is completed. When the unlabeled data are completely predicted as reliable samples, the entire iteration process is completed. Finally, all the samples are put into the classifier for training, and the final model obtained is the case-related classification model used in this study. The specific algorithm flow is as follows.

Algorithm 1. PU iterative algorithm pseudocode

```
Input: initial labeling data P; Unlabeled training set U; Test Set
         Classifier D;
         Topic model T;
Output: Probability y of News Involving News
Process:
1: t_model = T([U;P])
2: PN = [I-DNF(P,U);P]
3: U = U-PN
4: while len(U) > 0 do
5:     d_model = D(PN,t_model(PN))
6:     Probability of prediction = sort(d_model(U))
7:     for i = 1,2,...,2000 do
8:         if the prediction probability i is greater than the probability of being a positive sample do
9:             PN = PN + U[i]
10:            U = U - U[i]
11:        end if
12:        if the prediction probability-i is less than the probability of being a positive sample do
13:            PN = PN + U[-i]
14:            U = U - U[-i]
15:        end if
16:    end for
17: end while
18: d_model = D(PN,t_model(PN))
19: y = d_model(U)
```

4. Experiment

To evaluate the performance of our model, we conducted three experiments on a dataset that included case-related news. One was a comparative experiment that compared the performance of the PU classification algorithm without a topic. Simultaneously, we analyzed their prediction performance in iterative training. In addition, we conducted comparative experiments on initial datasets of different scales and an iterative step comparison experiment. Under different steps, we verified the effectiveness of our method and the PU classification algorithm without a topic. The experimental results also validated the effectiveness of our method on the relevance-analysis task of case-related news. This also illustrates that topic information enhances the PU learning iterative process and can improve the performance of the model.

4.1 Dataset

We use the categories provided in Table 1 to define the scope of case-related news. By crawling relevant news data from microblogs, Tianya forums, and other websites, we constructed a dataset. The length of the news text in this dataset was approximately 100 to 250 characters. To facilitate the experimental verification effect, we manually marked all the data, including 10,000 case-related and 20,000 non-related news. During the experiment, the marked data were regarded as unmarked data required for accurate analysis.

Table 1. Category of case-related news

Category of case-related news	Description	VisuShrink
Legal decision	The court officially pronounced the result of the case.	28.76
Report results of fighting black and anti-corruption	Announcement by the commission of discipline inspection and other state institutions on the phased results of legal tasks such as fighting black and anti-corruption.	26.46
Potential cases	The content of the report is news that involves the case or may develop into a case, such as tracking the progress of the case, the self-statement of the parties to the case, the description of the circumstances of the case, the vicious behaviors such as citizens' high-speed rail seats, drunk driving and escaping, etc.	25.14
Man-made accident	Accidents such as car accidents, man-made fires and other accidents for which people are responsible.	24.81

4.2 Parameter Setting and Evaluation Metrics

This study sets the maximum length of the body to 200 characters. The Adam algorithm is used as the optimizer. The learning rate is set to 0.001, dropout for single-layer BiLSTM is set to loss 0.2, batch processing size is set to 128, training rounds are set to 20, and the number of iterative trainings is the ratio of the total amount of unlabeled data to the number of positive and negative samples extracted each time. Our evaluation metrics mainly adopt accuracy (Acc), precision (P), recall (R), and F1. The accuracy describes the proportion of correctly predicted samples to the total number of samples. Precision describes the number of predicted positive samples that are truly positive. The recall indicates how many positive samples are predicted correctly, and the F1 value is the adjustment of accuracy and recall. In addition, we also use the error rate to analyze the verification results.

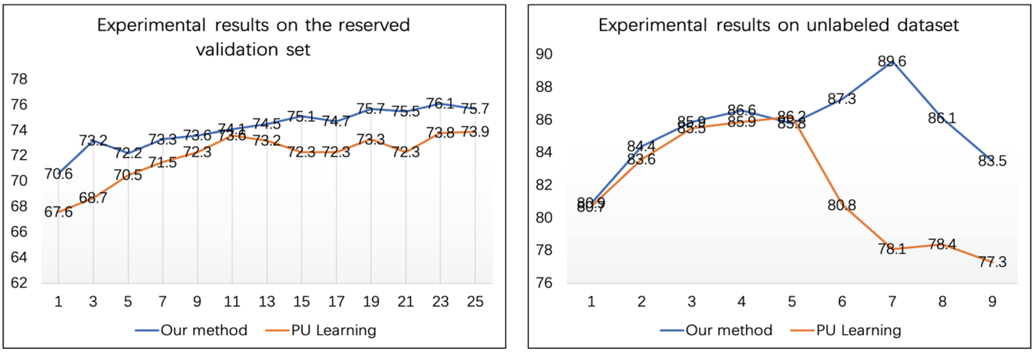
4.3 Experimental Results and Analysis

First, we compare our method with the PU learning baseline model using two groups of experiments and verify its effectiveness using a small amount of case-related news data. Subsequently, we compare our method with the advanced PU learning method, which also indicates that our method is competitive.

In addition, we also test some important parameters using two groups of experiments and determine that our method is superior to the baseline model of the PU method in each iteration when the initial data scale and iteration step are fixed; when the initial data scale is small or the iteration step is large, the performance will improve more stably.

4.3.1 Comparative experiment with PU learning baseline model

Because our method mainly improves the conventional two-stage PU learning, this experiment compares the performance of our method with the PU learning baseline model on the case-related news dataset. We established two experimental groups. One group used the reserved validation set to evaluate the generalization performance of the classifier trained in the iterative process and the other group evaluated the performance of the classifier trained in each iteration on the remaining unlabeled samples. The experimental results are illustrated in Fig. 4, where the x-axis represents the number of iterations, and the y-axis represents the size of the evaluation index. In the process of the experiment, the number of case-related news samples in the initial test was preset to 1000, the non-related news samples extracted were also 1000, and the experiment was compared with PU learning without topic and conventional classification model. Among them, "PU learning" refers to the PU learning method without a topic, and the classifier used is the same as that used in this study. In this experiment, the iteration step was set to 500, all the parameters of the two groups of experiments were the same, and the evaluation index was the F1 value.



(a) (b)

Fig. 4. The results of PU learning baseline and our method using different data: (a) illustrate adding the reserved validation data in training process and (b) illustrate adding the unlabeled data in training process.

We evaluated our method using the reserved validation set, and the results are illustrated in Fig. 4(a). As can be observed from the figure, the F1 value for PU learning was 73.9%. However, our method reached 75.7%, 1.8% more than PU learning. Fig. 4 illustrates that our method achieved effective results in filtering challenges with only a few case-related news samples. We can observe from the curve that the curve of our method is more stable and has better comprehensive performance than PU learning, and the classifier trained by each iteration has a certain degree of improvement inaccuracy. The data fully explained that the topic plays an enhanced role in PU learning. This improvement occurred not only in the first training but in every iterative process. It played an effective role in relieving "error accumulation" in PU learning.

Fig. 4(b) illustrates the evaluation results of our method of "unlabeled datasets" in the first nine iterations of training. These "unlabeled datasets" have been labeled manually and used as unlabeled data in the prediction process. As can be observed from the figure, the performance of our method in predicting unlabeled data is superior to conventional PU learning, and the gap between the two becomes larger as the number of iterations increases. In fact, in the first to fifth iterations, our method is only slightly improved compared with conventional PU learning, but in the subsequent iterations, with the increase in training data, the gap is gradually widened. For the seventh iteration, the F1 value of our method on the unlabeled dataset is approximately 11.5% ahead of the conventional PU learning. The reason for this phenomenon is that the method in this study uses the information of the text more effectively and realizes layer-by-layer enhancement of the conventional PU learning process.

4.3.2 Comparative experiment with advanced PU learning method

The main purpose of this experiment was to compare the performance of our method with that of other advanced PU learning methods. We chose the two latest PU learning methods for comparison. Among them, the nnPU model [24] is a classical model for PU learning. The idea was to change the weight of unlabeled samples by improving the loss function, and finally obtain the unbiased optimal solution. The I-PU model [25] calculated the probability according to the similarity between unlabeled samples and positive samples with the idea of ensemble learning, thus labeling unlabeled samples and generating multiple datasets to train different models. The experimental results are listed in Table 2.

Table 2. Comparative experiment with advanced PU learning method (unit: %)

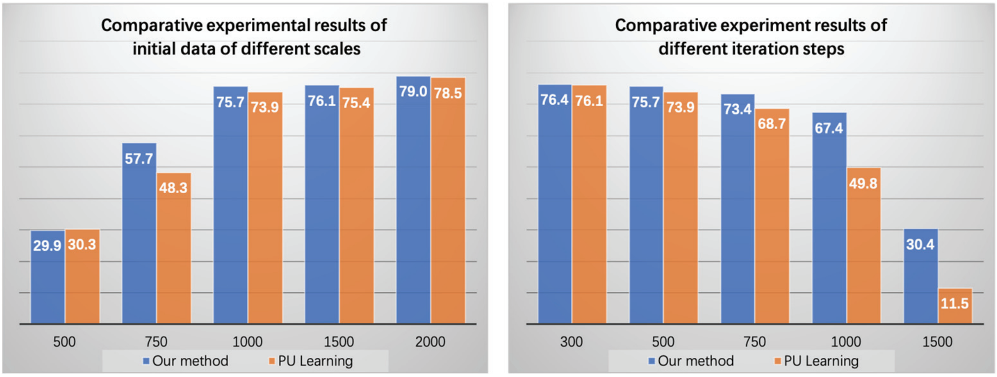
Evaluation metrics	Accuracy	Precision	Recall	F1
nnPU [24]	66.3	61.4	70.1	65.5
I-PU [25]	77.8	74.5	71.2	72.8
Our method	80.6	70.9	77.2	73.9

As presented in Table 2, the performance of our method for the case-related news dataset is better than that of the two advanced methods, of which the F1 value is 1.1% ahead of the I-PU model and 8.4% ahead of the nnPU model. On the one hand, because of the adaptability of the dataset, the case-related news dataset has the characteristics of a case-related topic. We extract the topic to strengthen the iterative process, which gives the classifier good domain characteristics. On the other hand, our method adopts an iterative training process, which is higher than the other two methods in terms of training duration and data utilization rate. It can be observed that our method is fairly competitive compared to other advanced methods in terms of performance.

4.3.3 Comparative experiment of different experimental parameters

The main purpose of this experiment was to observe the performance improvement of our method compared to conventional PU learning under different experimental parameters. We chose two important parameters to conduct the experiment: the initial data and iteration step sizes. The experimental results are illustrated in Fig. 5.

We compare our method with PU learning under different initial data scales, and the results are illustrated in Fig. 5(a). We set four different initial datasets, 500, 750, 1000, 1500, and 2000. The number of iterations is 500. The x-axis represents the different data scales. This figure presents the evaluation results obtained by iteration of the unlabeled data. When the initial data scale is merely 500, PU learning has already failed, as illustrated in Fig. 5(a). Failure occurs when PU learning depends on the scale of the initial labeled data. If the initial data scale is too small, the trained classifier lacks precision. The low precision results in reliable positive and negative samples with a large bias during the subsequent forecast process. With iteration processing, this bias accumulates and leads to PU learning failure. With an increase in the initial data scale, the bias of each iteration is smaller, and the final result is better, which is a common phenomenon in PU learning. Our method follows this phenomenon as well. Compared to PU learning, our method has better adaptability with small initial data. When the initial data size is only 750, the F1 value gap between the effect of our method and the conventional PU learning reaches 9.4%. As the initial data scale increases, the gap is smaller. This result indicates that our method is more effective in using messages on a small initial scale.



(a) (b)

Fig. 5. (a) The initial data of different scales and (b) the different iteration steps.

We compare our method with PU learning with different iteration steps, and the results are illustrated in Fig. 5(b). We set five different iteration steps: 300, 500, 750, 1000 and 1500. This figure presents the evaluation results obtained by iteration of the unlabeled data. First, we set the data scale to 1000, the x-axis represents different data scales. As illustrated in Fig. 5(b), the performance of our method and conventional PU learning is maintained at a good level. As the iteration step increases, the performance of PU learning will decrease, and our method will be kept well. Regardless of our method or PU learning, the classifier for each iteration of training has basic precision. This classifier predicts unlabeled samples and sorts them by probability. This leads to a higher density of reliable samples at both ends. In the low iteration step, both our method and PU's positive and negative samples can be classified well, which leads to a small bias of positive and negative samples. However, as the number of steps increases, the positive and negative sample biases of PU learning increase, and the performance decreases. Our method has an obvious fall when the step reaches 1000, which indicates that PU learning combined with topic has better adaptability. When the iteration step reaches 1500, our method fails, and PU learning also fails. When the initial data scale is 1000, the classifier trained by PU learning has a limited precision. Even if we add a topic for enhancement, the required precision cannot be achieved.

5. Conclusion

We proposed a filtering method for case-related news that combines topic and PU learning methods. The performance of the final classifier for PU learning strongly depended on the initial labeled data, in the case of a small amount of case-related news, the accuracy of case-related news filtering decreased. Therefore, we developed a topic-enhanced PU learning method to extract keywords from news data by adding unsupervised pre-training to ensure that the trained PU classifier had higher accuracy. Through repeated iterative training of the obtained classifier, the performance of the case-related news filtering model was improved. Thus, the performance of our method on the test set was 1.8% ahead of the F1 value of the PU learning baseline model and 1.1% better than that of the advanced PU learning baseline model. In the future, this method will continue to be optimized and applied to the case of public opinion analysis.

Acknowledgement

This study was supported by the project of the National Key Research and Development Project (No. 2018YFC0830100) and the Science and Technology Plan Projects of Yunnan province (No. 202001AT070046).

References

[1] Y. Shao, S. Taylor, N. Marshall, C. Morioka, and Q. Zeng-Treitler, “Clinical text classification with word embedding features vs. bag-of-words features,” in *Proceedings of 2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, 2018, pp. 2874-2878.

[2] A. Phung and M. Stamp, “Universal adversarial perturbations and image spam classifiers,” in *Malware Analysis Using Artificial Intelligence and Deep Learning*. Cham, Switzerland: Springer, 2021, pp. 633-651.

[3] R. Rahim, I. Zulkarnain, and H. Jaya, “A review: search visualization with Knuth Morris Pratt algorithm,” *IOP Conference Series: Materials Science and Engineering*, vol. 237, no. 1, article no. 012026, 2017. https://doi.org/10.1088/1757-899x/237/1/012026

[4] J. Song and C. Miao, “Optimization and implementation of Sunday algorithm,” in *Proceedings of 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, Dublin, Ireland, 2019, pp. 263-266.

[5] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, “A novel text mining approach based on TF-IDF and Support Vector Machine for news classification,” in *Proceedings of 2016 IEEE International Conference on Engineering and Technology (ICETECH)*, Coimbatore, India, 2016, pp. 112-116.

[6] S. M. Jang, T. Geng, J. Y. Q. Li, R. Xia, C. T. Huang, H. Kim, and J. Tang, “A computational approach for examining the roots and spreading patterns of fake news: evolution tree analysis,” *Computers in Human Behavior*, vol. 84, pp. 103-113, 2018.

[7] H. Yu, J. Han, and K. C. C. Chang, “PEBL: positive example based learning for web page classification using SVM,” in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002, pp. 239-248.

[8] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[9] S. Gururangan, T. Dang, D. Card, and N. A. Smith, “Variational pretraining for semi-supervised text classification,” in *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019, pp. 5880-5894.

[10] B. Liu, W. S. Lee, P. S. Yu, and X. Li, “Partially supervised classification of text documents,” in *Proceedings of the 19th International Conference (ICML)*, Sydney, Australia, 2002, pp. 387-394.

[11] L. M. Manevitz and M. Yousef, “One-class SVMs for document classification,” *Journal of Machine Learning Research*, vol. 2, pp. 139-154, 2001.

[12] Y. Ren, D. Ji, and H. Zhang, “Positive unlabeled learning for deceptive reviews detection,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 488-498.

[13] X. L. Li, P. S. Yu, B. Liu, and S. K. Ng, “Positive unlabeled learning for data stream classification,” in *Proceedings of the 2009 SIAM International Conference on Data Mining*, Sparks, NV, 2009, pp. 259-270.

[14] Y. Xiao, B. Liu, J. Yin, L. Cao, C. Zhang, and Z. Hao, “Similarity-based approach for positive and unlabelled learning,” in *Proceedings of 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 2011, pp. 1577-1582.

[15] J. J. Huang, P. W. Li, M. Peng, Q. Q. Xie, and C. Xu, “Review of deep learning-based topic mode,” *Chinese Journal of Computers*, vo. 43, no. 5, pp. 827-855, 2020.

[16] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proceedings of the 2nd International Conference on Learning Representations*, Banff, Canada, 2014.

[17] Y. Miao, L. Yu, and P. Blunsom, “Neural variational inference for text processing,” in *Proceedings of the 33nd International Conference on Machine Learning*, New York, NY, 2016, pp. 1727-1736.

[18] Y. Miao, E. Grefenstette, and P. Blunsom, “Discovering discrete latent topics with neural variational inference,” in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 2410-2419.

[19] R. Ding, R. Nallapati, and B. Xiang, “Coherence-aware neural topic modeling,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 830-836.

[20] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei, “Automatic differentiation variational inference,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 430-474, 2017.

[21] H. Yu, J. Han, and K. C. Chang, “PEBL: web page classification without negative examples,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 70-81, 2004.

[22] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, 2015.

[23] K. W. Church, “Word2Vec,” *Natural Language Engineering*, vol. 23, no. 1, pp. 155-162, 2017.

[24] R. Kiryo, G. Niu, M. C. D. Plessis, and M. Sugiyama, “Positive-unlabeled learning with non-negative risk estimator,” 2017 [Online]. Available: https://arxiv.org/abs/1703.00593.

[25] L. Jiang, D. Li, Q. Wang, S. Wang, and S. Wang, “Improving positive unlabeled learning: practical AUL estimation and new training method for extremely imbalanced data sets,” 2020 [Online]. Available: https://arxiv.org/abs/2004.09820.

Guanwen Wang **https://orcid.org/0000-0002-9201-6389**

She received B.S. degrees in School of Software from Zhengzhou University in 2019. And she is pursuing a master’s degree at the School of Information Engineering and Automation of Kunming University of Science and Technology, since September 2019. Her current research interests include natural language processing and Information retrieval, etc.

Zhengtao Yu **https://orcid.org/0000-0002-4012-461X**

He received the Ph.D. degree from Beijing Institute of Technology in 2005. He is a professor and Ph.D. supervisor at Kunming University of Science and Technology, and the director of Yunnan Key Laboratory of Artificial Intelligence. His research interests include natural language processing, machine translation, etc.

Yantuan Xian **https://orcid.org/0000-0001-6411-4734**

He received the M.S. degree from Shenyang Institute of Automation (SIA), Chinese Academy of Sciences in 2006. He is an associate professor at Kunming University of Science and Technology. He is a member of CCF. His research interests include natural language processing, information extraction, machine translation, etc.

Yu Zhang **https://orcid.org/0000-0003-2271-981X**

He received B.S. degrees of Chuzhou College in 2016. And he is pursuing a master's degree at the School of Information Engineering and Automation of Kunming University of Science and Technology, since September 2018. He current research interests include natural language processing, machine translation, etc.