**ABSTRACT**

As news reading on social network platforms becomes more and more popular, fake news becomes a social issue. The fake news can make use of social network platforms to mislead readers, which can arouse public anxiety and panic. The huge challenges are how to identify fake news and evaluate the user credibility of the news. Most of existing methods have shown excellent performance for fake news detection, since they tend to use a large amount of user information data to train a model. Moreover, it is extremely difficult to obtain a dataset with user information in real life. In order to address this problem, we take advantage of user correlations among news and propose a weak supervision fake news detection driven method to evaluate user credibility (WSEUC), which can be evaluated by the news of publishers published in the past. It consists of three parts: the automatic threshold producer, the collaborative credibility characterizer, and the dynamic credibility filter. Automatic threshold producer is responsible to extract textual features from posts and calculate the accuracy of the training set as the threshold of the testing set. The role of collaborative credibility characterizer is to calculate user credibility through past news and make it more authentic through collaborative filtering. It cooperates with dynamic credibility filter to eliminate the disturbance generated by the change in the training set and selects the high-quality unlabeled news to join training set. Experiments are conducted on user-related datasets collected from pilitifact.com and XXX. The experimental results show that WSEUC model can outperform state-of-the-art methods, and evaluate user credibility primely.

**1 INTRODUCTION**

With the development of Internet technology, social media represented by Weibo, WeChat, Twitter and Facebook have attracted a large number of users and gradually become an important medium for news creation and dissemination. Unfortunately, there are increasingly more kinds of fake news emerging on social media constantly. On the one hand, the emergence of fake news is often accompanied with hot-spot events on social media, which is easy to arouse the emotional resonance of the public and form hot-spot topics. On the other hand, verifying the truth of news needs a high threshold of professional knowledge, which is not difficult to confuse the public and deceive the public. A number of fake news with exaggerated claims or lack of factual evidence spreading in social media will often cause misunderstanding and panic among consumers, which damaging the authority of media and government. Therefore, accurate fake news detection in social media is of great significance for halting the spread of fake news.

So far, there are a variety of approaches of detecting fake news including traditional machine learning models and deep learning models. Traditional machine learning models usually extract features from news and train classification model. Compared with the traditional learning methods, the existing deep learning models have better performance in detecting fake news due to their strong automatic feature extraction ability. However, they are still not able to deal with the specific challenge of fake news detection, i.e., evaluating user credibility on existing publishers. Owing to the existing models require plenty of user information data that are sometimes difficult to obtain and cost high. Meanwhile, training a deep learning model often requires a mass of hand-marked data. In addition, accurate tags can only be obtained if the annotators have sufficient knowledge of current affairs. The production of such labeled data is expensive and time-consuming. For this reason, instead of using a mass of hand-marked data, we believe that mining the hidden information behind a small amount of limited labeled data can effectively improve the ability of deep learning model.

To mine the hidden information behind a small number of limited labeled data, the first step is to discover the publisher information behind the news. For different publishers, they have their own distinctive features that are not measurable. We try to develop a solution to evaluate user features from the new perspective of user correlations among news. The existing researches of fake news detection usually ignore the relationship between different news published by the same user. However, a user may publish multiple news with similar content in the social network. It is a technically problem to evaluate user features and mine relationship between different news published by the same user. First, there are various attributes such as user age and education level evaluating user features. However, these data are difficult to obtain and the authenticity of these data are difficult to guarantee. Second, the relationship between different news published by the same user is invisible and immeasurable. Thus, how to effectively use user correlations among news to improve the accuracy of fake news detection and evaluate user without user attributes data are the challenges that we have to address.

In order to address these challenges, we propose a weak supervision fake news detection driven method to evaluate user credibility (WSEUC). Since news is distributed by different users, we propose a user evaluation criterion - user credibility. We use the credibility of their news published in the past to evaluate the user credibility, which outperforms the existing models. We utilize the collaborative filtering based on user to mine the similarity among users, and use the adaptive Kalman filter to reduce the fluctuation of user credibility to enhance its stability. Then, by using the weak supervised learning method, the small-scale labeled data are used to mark the large-scale unlabeled data, which are incorporated into the training set. It consists of three components: the automatic threshold producer, the collaborative credibility characterizer, and the dynamic credibility filter. The automatic threshold producer is responsible to extract textual features and produce an evaluation criterion for the authenticity of unlabeled data. The role of collaborative credibility characterizer is to calculate user credibility by the news that the user published in the past and make it more authentic through collaborative filtering. It cooperates with dynamic credibility filter to eliminate the disturbance generated by the change of training set in each iteration. It takes advantage of the textual characteristics and user credibility to mark the unlabeled data and choose high-quality data into training data, thus expanding the training set and improve the detection performance. Experiments are performed on user-related datasets collected from pilitifact.com. The experimental results show that WSEUC model can be superior to state-of-the-art models, and evaluate user appropriately.

The main contributions of this paper are as follows:

As far as I know, we are the first propose a user evaluation criterion-user credibility, which can evaluate user credibility by the news that user published, and can help improve the accuracy of fake news detection.

The proposed WSEUC uses collaborative credibility characterizer to measure the user credibility among different users, and further learns the user features which can promote to new users.

We propose WSEUC to automatically mark unlabeled news, which expands the size of the training set and improves the results of our model.

Our experimental results prove that WSEUC can evaluate user credibility appropriately and effectively identify fake news. At the same time our model can be superior to state-of-the-art models on two large scale datasets.

The rest of paper is organized as follows: research in the related field is summarized in Section 2, the details of our model are introduced in Section 3, experimental results are shown in Section 4, and the study is concluded in Section 5.

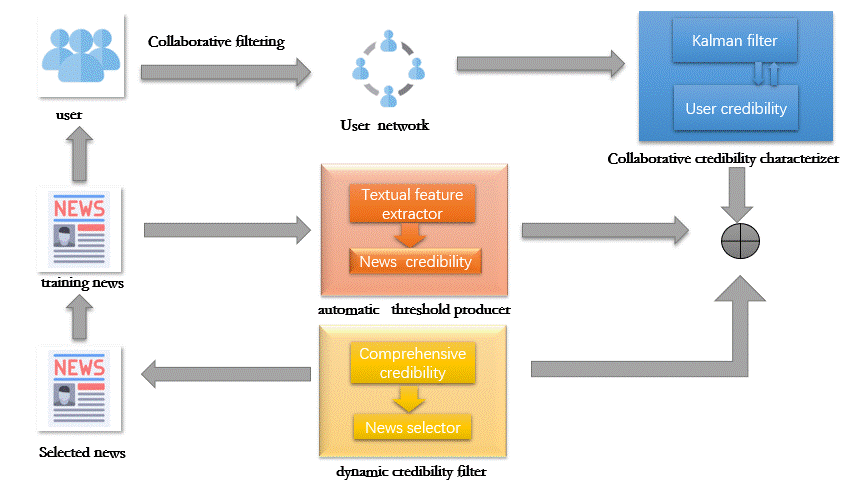
**2 RELATED WORK**

**3 METHODLOGY**

To evaluate user credibility appropriately without user information and improve the detection performance of fake news on small-scale dataset, we propose a weak supervision fake news detection driven method to evaluate user credibility. We first introduce the three components: automatic threshold producer, collaborative credibility characterizer, and dynamic credibility filter. Then describe how to integrate three components to calculate user credibility and identify fake news. The details of the algorithm flow are shown in the last subsection.

**3.1 Model Overview**

The goal of our model is to calculate appropriate user credibility to evaluate publisher credibility and improve the detection performance of fake news. As shown in Fig 1, our proposed method WSEUC integrates three components: the automatic threshold producer, the collaborative credibility characterizer, and the dynamic credibility filter. First of all, the automatic threshold employs a word embedding layer and Text-CNN to extract textual features and position features from input news, and produce an evaluation criterion for the authenticity of unlabeled data. Based on the automatic threshold producer, the collaborative credibility characterizer calculates user credibility by the news that the user published news in the past, and employs collaborative filtering of user by limited Boltzmann machine which discovers similarities between users. After that, the dynamic credibility filter employs Kalman filter to eliminate the disturbance generated by the change in the training set and select the high-quality unlabeled data to join training set.



**Fig.1. The framework of WSEUC**

**3.2** **automatic threshold producer**

The sequential list of words in the news is the input to the news characterizer. For detailed procedures of the textual feature extractor, each word in the text is represented as a word embedding. The word embedding for each word is initialized with the pre-trained word embedding on the given dataset. For -th word in the news, its corresponding dimensional word vector can be represented as . Thus, a sentence with words can be represented as:

**图片包含 游戏机

描述已自动生成**

**Fig.2.** **The architecture of Text-CNN.**

(1)

Where is the concatenation operator. A convolutional filter with window size selects consecutive words in the news as inputs and outputs one feature. In order to clearly demonstrate the process, we select consecutivewords starting with the -th word for example, the filter operation can be represented as:

(2)

where represents the weight of the filter and is the ReLU activation function. Next, we apply this filter to the rest of words of the news and get a feature vector for this news:

(3)

For every feature vector, we use max-pooling operation to select the largest feature value to extract the most important feature. Similarly, we use filters with multiple window sizes to extract textual feature at different granularities. For a specific window size, we have different filters. Therefore, assuming there are possible window sizes, we have filters in total. The textual feature representation after the max-pooling operation can be written as , which is the output of the news characterizer.

Then, we introduce the fake news detector. It deploys a fully connected layer with softmax to output the news credibility. The fake news detector takes textual feature representation as input. We denote the fake news detector as , where represents all the parameters included. The output of the fake news detector is represented as news credibility, denoted as .

In the fake news detector, our labeled news set is used to train a LSTM network. We denoted this network as , where represents all the parameter included. First, the labeled news set is fed into textual feature extractor to get news textual feature represented as . Then, is fed into LSTM network to obtain news credibility, denoted as . Afterwards, the -th labeled news textual feature, denoted as , the news credibility , can be represented as:

(4)

The news credibility set is , where is the amount of labeled news. We use to present the labels of the labeled news set and to present the predicted labels of the labeled news set. Then we employ cross entropy to calculate the detection loss:

(5)

We minimize the loss function by seeking the optimal parameters , and this process can be represented as:

(6)

We compare the predicted labels set with the real labels to calculate the accuracy value. The accuracy value is the threshold represented as that we select to represent the evaluation criterion for the unlabeled news. Finally, the output of the automatic threshold producer can be written as and .

As previously discussed, one of the major challenges for fake news detection is how to distinguish whether the users like to publish real news. This requires us to be able to evaluate user credibility accurately. Direct minimization of detection loss only can help us obtain the credibility of news content. However, the user credibility can-not obtain well. thus, we need to make the model learn the user features, which can be evaluated as user credibility. To accomplish the goal, we need to find an evaluation method to calculate the user credibility of each news. In particular, we use the overall reliability of the news posted by users in the past to represent the user credibility.

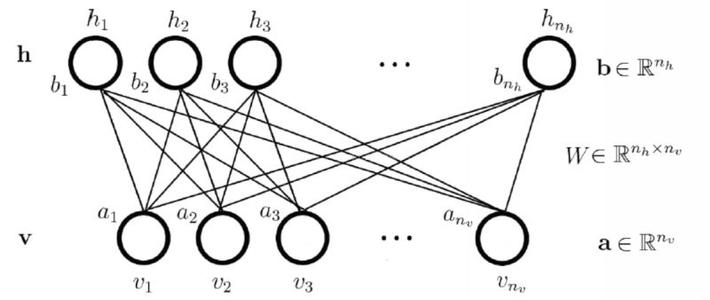
**3.3 collaborative credibility characterizer**

For -th labeled news, we can obtain the user credibility . The entire user credibility set is , where is the number of the labeled news. In the labeled news set, the -th news is true or false is represented as where =1 means true and =0 means false. For different users, is represented as the sum of the news that the user published in the labeled news set. The process of calculating user credibility can be represented as:

(7)

In order to explore the connections between different users and make user credibility more authentic, we amend user credibility further. First, the news with similar content is clustered into categories. we make all the news for each user into a certain category, and the category is denoted as (=1,2,3...). For -th user and -th category, is represented as user ratings. The user-news matrix is represented as:

(8) The collaborative credibility characterizer employs restricted Boltzmann machine (RBM) to perform collaborative filtering. The RBM use the user-news matrix with integer ratings between and where is represented as the max of the ratings, and 0 when a rating is miss. It has a linear visible layer of units , which is fully connected with a hidden layer of units H=. There are some special weight units: the bias for the visible layer , and the bias for the hidden layer . All the connection weights between visible layer and hidden layer are the symmetric.The structure of restricted Boltzmann machine is as Fig2.



**Fig3 the structure of restricted Boltzmann machine**

The activation function is logistic function, and the activation function of hidden layer conditioned on the visible layer is as follows:

(9)

(10)

In brief, during learning process of our RBM, it is trained to approximate the distribution by observed ratings distribution. To update the weights that are connected visible units and hidden units, we make RBM learn the weights as follows:

(11)

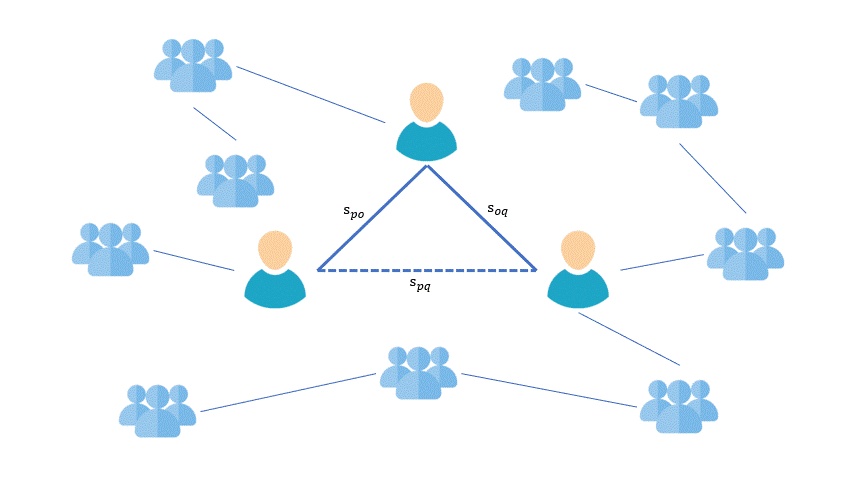
(12)

(13)

The predicted ratings predict by RBM model is new user-news matrix. To find the similarity between different users, we calculate the user similarity value represented as by Pearson correlation. The process of calculating is as follows:

(14)

Here and are represented different users. In order to mine the user similarity further, we build a user relationship network. In the user relationship network, the different nodes represent different users, and the edges represent the user similarity value between different users. If there is an edge between two users, we regard them as neighbors. If there is a common neighbor between two unassociated users and , they will be regarded as generation neighbors whose user similarity is represented as:

(15)

**Fig 4 the common neighbor in user relationship network**

Subsequently, we take user credibility through the user similarity between users in the same category and take the weighted average. The weighted user credibility is represented as whose process of calculating is as follows:

(16)

Finally, the output of the collaborative credibility characterizer can be written as .

**3.4** **dynamic credibility filter**

In the subsection, we introduce the dynamic credibility filter. It deploys Kalman filter to reduce the fluctuates of user credibility and improve detection performance of fake news. Kalman filter is an algorithm that uses linear state to estimate the optimal system state through observation data. The dynamic credibility filter is built on the top of the collaborative credibility characterizer, and takes observed user credibility generated from the collaborative credibility characterizer as input. We assume current iteration as and the last iteration as . We use predicted user credibility and observed user credibility to obtain optimized user credibility represented as follows:

  (17)

Then, we assume that is the predicted noise, which is usually assumed to follow Gauss distribution. and  denotes the corresponding noise covariance. We use optimized user credibility to predict optimized user credibility , and the process as follows:

(18)

is the corresponding observation noise. It is similar to , where and denotes the corresponding noise covariance. Then, we use observed optimized user credibility to predict observed user credibility

(19)

We use to denote optimized error covariance. Subsequently, the observed user credibility is the value of predicted user credibility. The predicted error covariance and predicted user credibility that we predict are as follows:

(20)

(21)

To update user credibility, we define Kalman gain is represented as:

(22)

Here H is a transfer matrix, which coordinate observed user credibility and predicted credibility. Finally, we obtain the optimized user credibility and optimized error represented as:

(23)

(24)

In the next iteration , and  are the initial value of the next iteration. The process of update and is as above. In this iteration, the ultimate user credibility is .

The entire news credibility set from automatic threshold producer is , and the entire user credibility set is . we use set as weight set. For each weight represented as in , we denote comprehensive credibility of news as whose calculating process is as follows:

(25)

Subsequently, we choose the maximum and minimum of comprehensive credibility represented as and in comprehensive credibility set . Then, the normalization of the comprehensive credibility is as follows:

(26)

We compare each unlabeled news with the from automatic threshold producer. If the comprehensive credibility , we assume the news as true, which is equal to label the news “1”. Otherwise we assume it as false, which is equal to label the news “0”.

Due to pseudo labels is often noisy, adding the unlabeled news with the pseudo label will reduce detection performance. To address this problem, we choose the high-quality unlabeled news based on comprehensive credibility. we sort the unlabeled news by comprehensive credibility. Subsequently, the unlabeled news ranked in the top five percent and bottom five percent that represented as the truest news and the most false news will be selected to join the training set for the next iteration. Therefore, selecting high-quality news can improve the quality of training set and expand the number of training set in the next iteration.

**3.5 model integration**

During the training stage, we feed the training set into the automatic threshold producer to obtain the informative textual feature representations . Subsequently, we pretrain the fake news detector using the training set . Again, we feed the training dataset into the fake news detector to find the threshold for unlabeled news. After that, we feed the unlabeled news set into the fake news detector to get the textual feature representations , and input into the fake news detector to assign pseudo labels to the unlabeled news set . The collaborative credibility characterizer calculates the user credibility set .The proposed dynamic credibility filter will select high-quality samples from the pseudo labeled dataset . Then both the selected news set, denoted as , and the original labeled news set are combined to update the training set. Thus, the final loss of the WSEUC is represented as:

(27)

Here, and are the losses on a small number of manually labeled news and automatically assigned news set separately, and controls the balance between them. We simply set the value of and as 1. The two losses are defined by the cross entropy as:

(28)

(29)

The detailed steps of proposed method are summarized in Algorithm.

|  |
| --- |
| **Algorithm** A weak supervision fake news detection driven method to evaluate user credibility |
| **Input:** Labeled data set , ground truth label setunlabeled data set , the number of ground true new for each user, the sum of the news for each user |
| **Output:** pseudo label set, comprehensive credibility set |
| 1: **for** number of training iterations **do**  2: Feed the labeled news set into news characterizer to obtain textual feature representations .  3: Train the fake news detector using the training set  4: Find the threshold t by again feeding the training set into the automatic threshold producer.  5: Feed the unlabeled news set into the fake news detector to obtain textual feature representations .  6: Feed into the automatic threshold producer to obtain the news credibility set .  7: Feed and into the collaborative credibility characterizer to obtain the user credibility set .  8: Feed into the collaborative credibility characterizer to collaborative filtering of users and update the user credibility set .  9 Feed the user credibility set into dynamic credibility filter and get the user credibility by Kalman filter.  10 Combine the user credibility set and news credibility set into comprehensive credibility set and select high-quality news samples .  11: Combine the selected news samples and the original labeled news set to update the training set. |
| 12: **end for** |

4. Experiments

In this section, we describe our experimental details. In Section 4.1, we first introduce our two real-world datasets. In Section 4.2, we state the baselines and the settings of baselines parameters. Section 4.3 clarifies the implementation details of our proposed WSEUC. Subsequently, we compare WSEUC mode with seven detection models to verify the performance of our model WSEUC and analyze the performance of all the baselines in Section 4.4. Finally, we analyze the performance of our model. In Section 4.5, we investigate its settings of the weight a to balance the news credibility and the user credibility and analyse the best weight a. To verify the robustness of our model, we analyse classification of the pseudo label in each iteration in Section 4.6. In Section 4.7, we investigate its setting of the iteration t and analyse the performance our model in each iteration.

4.1 Dataset

To fairly evaluate the performance of the WDEUC model, we conduct the experiments on two large real-world datasets, which are collected from political.com and Twitter. Then, we provide the details of the two datasets.

**LIAR**

The LIAR dataset is verified by fact-checking PolitiFact, which is used in[] for fake news detection. We divide the dataset into two parts that twenty percent of the dataset is training set and eighty percent is testing set. And the information in the dataset contains the writers of each news. In this work, we focus on detecting fake news by incorporating both text and the credibility of writer. Thus, we remove the writer with only a piece of news, and retain users who post more than or equal to two pieces of news.

We use the ground truth label collected from famous social website from Twitter[].

We divide the dataset into two parts that twenty percent of the dataset is training set and eighty percent is testing set. Following tradtional

|  |  |  |
| --- | --- | --- |
|  | The number of news | The number of users |
| Labeled news | 1684 | 614 |
| Unlabeled news | 6735 | 452 |

**Table 1 the Lair dataset**

4.2 Baselines

To compare the performance of WDEUC model with the relevant models, we choose baselines from the following three categories: traditional machine learning models, deep learning models, and the variant of the proposed model.

**traditional machine learning models**

**SVM.** We use the normalized textual feature representations and the ground truth label set to train a SVM model. The parameter kernel function is set as *RBF* and *C* is also set as 50.

**LR.**We use the normalized textual feature representations and the ground truth label set to train a Liner Regress model. The parameter *solver* is set as *lbfgs.*

**RF.** We use the normalized textual feature representations and the ground truth label set to train a Random Forest model. The parameter *n\_estimators* is set as 50.

**Deep learning model**

**LSTM.** LSTM employs one-layer as text feature extractor, the sentence embeddings come form news characterizer. The hidden nueral units are 128, and then the fully connected layer take the features as input to output the probability of the true news.

**EANN.**EANN is one of state-of-the-art model which contains three parts:feature extractor, event discriminator and fake news detector. Without image information in our dataset , we remove the image feature extractor and use textual extractor to detect fake news. The details of setting are the same in [] .

**MVAE.**

**Variant of the proposed model**

in this proposed model, we leverage both text and user credibility to detect fake news.

We remove the collaborative credibility characterizer and dynamic credibility filter separately and name them as WSECand WSEU. At the same time, we ingore the user correlations of news, and remove the user credibility in our model which named WSEUC.

**WSEU.** In order to verify the effect of the collaborative credibility characterizer in our model, we design a variant model, named WSEU.The WSEU contains three components: automatic threshold producer, collaborative credibility characterizer, and high-quality news selector.

**WSEC.** In order to verify the effect of the dynamic credibility filter in our model, we design a variant model, named WSEC.The WSEC contains three components: automatic threshold producer, collaborative credibility characterizer, and high-quality news selector.

**WSEUC-.** In order to verify the effect of the user in our model, we design a variant model, named WSEUC. The WSEUC contains: automatic threshold producer, high-quality news selector. It remove collaborative credibility characterizer and dynamic credibility filter. It ignore the user correlation of news, and only select high-quality news to training set by news features.

4.3 Implementation Details

In the automatic threshold producer, we set k=32 for dimensions of word-embedding. We set h=10, and the window size of filters varies from 2 to 5 in Text-CNN. The size of the last fully connected layer is 10. In the fake news detector, we set the hidden size of neural units is respectively 128. We set Adam as the optimizer, learning rate as 0.001, batch\_size as 50, and training epochs as 50. In the collaborative credibility characterizer, we set the category in user-news matrix as 20. in the dynamic credibility filter,we set the initial value as 0.02, as 0.01, as 1.0, as 1.0 andas 0.01. in all experiments, we employ Keras 2.3 and Python 3.7.

4.4 Performance Comparison

In the experiments, we conduct *SVM,LR,RF,LSTM,EANN,MVAE,WSEUC* model. Table 3 shows the experimental results of baselines. Then, we can observe the performance of all the models.

Among the seven detection methods, WSEUC always achieve the highest detection accuracy on Liar dataset and highest all the metrics on Twitter.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Method | Accuracy | AUC—ROC | Precision |  | Recall |
| Liar  Twitter | SVM | 0.5574 | 0.5438 | 0.5899 | 0.6258 | 0.6664 |
| LR | 0.5713 | 0.5439 | 0.5843 | 0.6722 | 0.7912 |
| RF | 0.5641 | 0.5472 | 0.5909 | 0.6405 | 0.6993 |
| LSTM | 0.5647 | 0.5223 | 0.5679 | 0.6977 | 0.9045 |
| UCDFM | 0.6085 | 0.5928 | 0.6259 | 0.6755 | 0.7327 |
| EANN  MVAE  SVM  LR  RF  LSTM  UCDFM  EANN  MVAE | 0.5986  0.5567  0.7867  0.7497  0.7888  0.7097  0.9322  0.7807 | 0.6165  0.5243  0.7856  0.7448  0.7838  0.6964  0.9286  0.7758 | 0.6178  0.5547  0.7661  0.7508  0.7990  0.7285  0.9659  0.7720 | 0.8086  0.6915  0.7687  0.7143  0.7585  0.6363  0.9229  0.7486 | 0.6974  0.9177  0.7714  0.6813  0.7219  0.5648  0.8837  0.7267 |

4.5 Parameters Analysis

in this section, we set different values of the weight a which balances the news credibility and user credibility. Along with the changes the values a ,the changes of the performance on two datasets are shown in Fig4.

4.6 Class Distribution Analysis

To analyze the pseudo labels assigning process of our model, we record the dynamic class distribution that the pseudo labels of unlabeled news in each epoch as Fig.

4.7 Epoch Analysis

In order to verify the effect of the high-quality unlabeled news we select,