Authorship Attribution for Short Texts with Author-Document Topic Model

Haowen Zhang, Peng Nie, Yanlong Wen, and Xiaojie Yuan(✉)

College of Computer and Control Engineering, Nankai University, Tianjin, China
{zhanghaowen,niepeng,wenyanlong,yuanxiaojie}@dbis.nankai.edu.cn

Abstract. The goal of authorship attribution is to assign the controversial texts to the known authors correctly. With the development of social media services, authorship attribution for short texts becomes very necessary. In the earlier works, topic models, such as the Latent Dirichlet Allocation (LDA), have been used to find latent semantic features of authors and achieve better performance on authorship attribution. However, most of them focus on authorship attribution for long texts. In this paper, we propose a novel model named Author-Document Topic Model (ADT) which builds the model for the corpus both at the author level and the document level to figure out the problem of authorship attribution for short texts. Also, we propose a new classification algorithm to calculate the similarity between texts for finding the authors of the anonymous texts. Experimental results on two public datasets validate the effectiveness of our proposed method.

Keywords: Authorship attribution · Topic model · Short text

1 Introduction

Authorship attribution has attracted much attention over the last decades because of its important role in criminal law, military intelligence, and humanities research [1]. The most majority of such researches try to determine the authors of controversial long texts, such as books, papers and so on. A lot of statistical learning methods have been used for authorship attribution and achieve good performance. In the recent years, more content is in the form of the short message with the growing popularity of Internet-based communication facilities, which creates great interest both in theory and computation on short texts. Compared to long texts, short texts have a few words, unclear structure and irregular usage of words. These characteristics make authorship attribution for short texts difficult, and the approaches that work well on long texts cannot give the same performance on short texts. As a result, finding the author of short texts (e.g., emails [2], blogs [3], twitters [4]) has attracted many researchers.

Support Vector Machine (SVM) is widely used in text classification. However, when it comes to authorship attribution, it cannot achieve the same performance because the goal of authorship attribution is not only to find texts which are

W. Liu et al. (Eds.): KSEM 2018, LNAI 11061, pp. 29–41, 2018.
https://doi.org/10.1007/978-3-319-99365-2_3

similar in content but also to consider the authors' writing style behind the content.

Finding the semantic feature of the author behind the content of the texts is very important for authorship attribution. Conventional topic models, such as PLSA [5] and LDA [6] assume that a document is a mixture of topics, where a set of correlated words is considered to be selected from the same topic. After enough number of iterations for training the model, words that appear in the same issue will be more likely assigned to the same topic. Based on this idea, LDA-H [7] first uses LDA to address the problem of authorship attribution and achieves better performance than SVM.

With the development of social media, the sparsity of content in short texts brings a new challenge to authorship attribution. To solve it, most of the early works [3,4] prefer to aggregate short texts into a lengthy pseudo-document and build a feature set for each author. However, they ignore the effect of each text for authorship attribution in this way.

In this paper, we propose a novel topic model named Author-Document Topic Model (ADT) for authorship attribution on short texts. We combine two level topic models Author-Topic Model (AT) and traditional LDA as the final model, which treats every word of each document equally at both two levels. Our motivation is based on the observation that when an author writes a document, the word frequency of a document varies apparently with authors of different writing style, and documents talked about different things will also affect the usage of words. To find the most probable author of the anonymous document, we propose a new classification algorithm to calculate the similarity between the training documents and the given document. Then the author of the most similar document will be assigned to the target document.
The main contributions of this paper include:

(1) We propose a novel generative model ADT for authorship attribution on short texts. We train ADT at both author level and document level. On the one hand, we use ADT to deal with the sparsity of content in short texts by aggregating short texts into lengthy pseudo-document. On the other hand, we use traditional LDA to make full use of training texts to find the authors of the anonymous texts.
(2) We propose a new classification algorithm for authorship attribution, which combines the influence of both authors and documents to calculate the similarity between the training document and the anonymous document.
(3) Experiment evaluations on two real-world datasets Pan'11 [8] and Blog [9] demonstrate the effectiveness of our proposed method. Compared to the current state of the art, ADT obtains a 6.63% improvement on Pan'11 and a 7.66% improvement on Blog.

The remainder of this paper is organized as follows. In the next section, we introduce the related studies in authorship attribution. In Sect. 3, we propose our ADT model and classification algorithm. The experimental evaluation is described in Sect. 4. In Sect. 5, we draw our conclusions.

2 Related Work

Authorship attribution is a traditional problem which can date back to the end of 19th century. It can be divided into two categories: similarity-based methods and machine-learning-based methods [1].

In similarity-based methods, SCAP [10] is the simplest method which calculates the Jaccard similarity between a given text and the profile texts of authors to find the most similar author. The feature sampling method (FS) [3] thinks that they do not know which features of the corpus are important for authorship attribution and which are not, so they randomly choose certain features from the feature set every time for calculating the similarity between the anonymous text and all authors' profiles. After repeating this process k times, the anonymous text is assigned to the author whose profile is most similar to the given text for a certain fixed number of the k times.

In machine-learning-based methods, SVM and topic model achieve better performance than others. SVM is often used for authorship attribution for its proven effectiveness in text classification and stability in handling many features [11,12]. Schwartz et al. [13] select features with k-signature, then they combine the feature set with flexible patterns for distinguishing authors, which is applied for SVM.

However, SVM tends to assign texts with similar features to the same author, which we think is not enough for authorship attribution. To obtain the latent semantic features of authors, the topic model LDA is applied to authorship attribution, and they use the Hellinger distance for calculating the similarity between document topic distribution to get the most similar author [7]. And two disjoint topic sets (DADT) [14] are trained separately, and they obtain better performance than ever before. Recent researches [4,15] focus on authorship attribution with the authors whose writing style is changed over time.

To the best of our knowledge, most of the previous works [3,4,7,13] only focus on the semantic features of authors, and they usually aggregate short texts into a lengthy pseudo-document and build feature sets for authors. In this paper, we are inspired by DADT [14] which builds the model for the corpus at both author level and document level and propose ADT model. Our ADT makes a great improvement on DADT, and we achieve better performance on authorship attribution for short texts.

3 Author-Document Topic Model

In this section, we detail the proposed model ADT. Firstly, we give the detail of ADT. Secondly, we explain the model inference. Thirdly, we describe the classification algorithm to find the most similar authors of the given texts. Finally, we make a comparison between ADT and DADT.

3.1 Topic Extraction via ADT

The graphical representation of ADT is shown in Fig. 1. It can be divided into two parts: Author-Topic Model on the left and Document-Topic Model on the right.

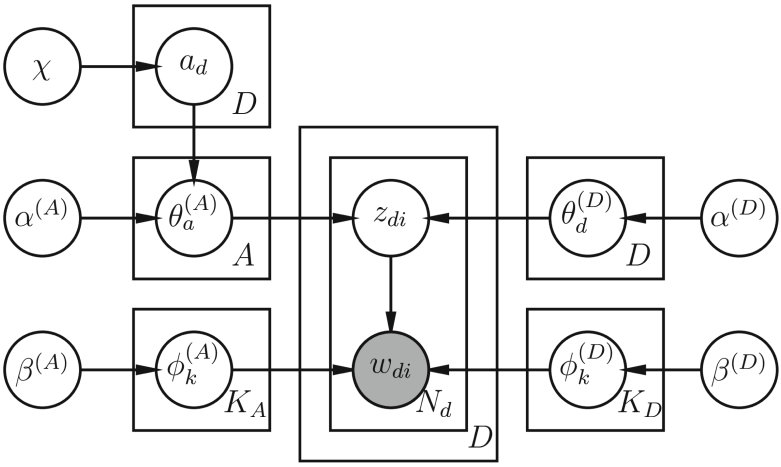

Fig. 1. The Author-Document Topic Model

We assume that the corpus has A authors, D documents, and V unique words in the vocabulary. $\mathcal{D}$ and $\mathcal{C}$ are used as the representation of the Dirichlet and categorical distributions respectively. Besides, K_A and K_D denote the number of topics of author level and document level. Formally, ADT assumes the following generative process for each document in a corpus **D**:

Author Level

1. Draw an author distribution χ, which is determined by the number of documents for each author;
2. Draw an author topic distribution $\theta_a^{(A)} \sim \mathcal{D}(\alpha^{(A)})$ for each author a;
3. Draw a word distribution $\phi_k^{(A)} \sim \mathcal{D}(\beta^{(A)})$ for each author topic k;
4. Draw a word distribution $\phi_k^{(D)} \sim \mathcal{D}(\beta^{(D)})$ for each document topic k.

Document Level

1. Draw a document topic distribution $\theta_d^{(D)} \sim \mathcal{D}(\alpha^{(D)})$ for each document;
2. Draw document's author $a_d \sim \mathcal{C}(\chi)$.

Word Level

For each word i in document d:

1. Draw an author topic $z_{di}^{(A)} \sim \mathcal{C}(\theta_{a_d}^{(A)})$ and word $w_{di} \sim \mathcal{C}(\phi_{z_{di}}^{(A)})$;
2. Draw a document topic $z_{di}^{(D)} \sim \mathcal{C}(\theta_d^{(D)})$ and word $w_{di} \sim \mathcal{C}(\phi_{z_{di}}^{(D)})$.

3.2 Model Inference

Given the parameter $\theta_a^{(A)}$, $\theta_d^{(D)}$, $\phi_k^{(A)}$ and $\phi_k^{(D)}$, the conditional probability of word w_i is estimated as follows:
Author level:

$$\begin{aligned} p(w_i|\theta_a^{(A)}, \phi^{(A)}) &= \sum_{k=1}^{K_A} P(w_i, z_i = k|\theta_a^{(A)}, \phi^{(A)}) \\ &= \sum_{k=1}^{K_A} P(z_i = k|\theta_{ak}^{(A)}) P(w_i|z_i = k, \phi_{k,w_i}^{(A)}) \\ &= \sum_{k=1}^{K_A} \theta_{ak}^{(A)} \phi_{k,w_i}^{(A)}. \end{aligned} \tag{1}$$

Document level:

$$\begin{aligned} p(w_i|\theta_d^{(D)}, \phi^{(D)}) &= \sum_{k=1}^{K_D} P(w_i, z_i = k|\theta_d^{(D)}, \phi^{(D)}) \\ &= \sum_{k=1}^{K_D} P(z_i = k|\theta_{dk}^{(D)}) P(w_i|z_i = k, \phi_{k,w_i}^{(D)}) \\ &= \sum_{k=1}^{K_D} \theta_{dk} \phi_{k,w_i}. \end{aligned} \tag{2}$$

For all words, ADT maximizes the likelihood function for the corpus in two levels:
Author level:

$$p(\mathbf{D}|\theta_a^{(A)}, \phi^{(A)}) = \prod_{d=1}^{D} \prod_{n=1}^{Nd} \sum_{k=1}^{K_A} \theta_{ak}^{(A)} \phi_{k,w_i}^{(A)}. \tag{3}$$

Document level:

$$p(\mathbf{D}|\theta_d^{(D)}, \phi^{(D)}) = \prod_{d=1}^{D} \prod_{n=1}^{Nd} \sum_{k=1}^{K_D} \theta_{dk} \phi_{k,w_i}. \tag{4}$$

There are two common ways variational inference [6] and collapsed Gibbs sampling [16] for inferring topic models. We use collapsed Gibbs sampling to conduct approximate inference for $\theta_a^{(A)}$, $\theta_d^{(D)}$, $\phi_k^{(A)}$ and $\phi_k^{(D)}$. For i_{th} word of d_{th} document, its author is known as author a, ADT samples author topic $z_i^{(A)}$ according to the following conditional distribution:

$$p(z_i^{(A)} = k, x_{di} = a|z_{-i}^{A}, \mathbf{D}) \propto \frac{(n_{-i,k|a} + \alpha^{(A)})}{(n_{-i,*|a} + K_A\alpha^{(A)})} \frac{(n_{-i,w_i|k}^{(A)} + \beta^{(A)})}{(n_{-i,*|k}^{(A)} + V\beta^{(A)})}. \tag{5}$$

where $z_{-i}^{(A)}$ is the topic assignments for all words, $n_{-i,k|a}$ is the number of words assigned to author topic k in author a, $n_{-i,*|a}$ is the total number of words in author a, $n_{-i,w_i|k}^{(A)}$ is the number of word w_i in author topic k, and $n_{-i,*|k}^{(A)}$ is the total number of words in author topic k. All of them exclude current assignment of $z_i^{(A)}$.

In the same way, ADT sample document topic $z_i^{(D)}$ according to the following conditional distribution:

$$p(z_i^{(D)} = k|z_{-i}^{D}, \mathbf{D}) \propto \frac{(n_{-i,k|d} + \alpha^{(D)})}{(n_{-i,*|d} + K_D\alpha^{(D)})} \frac{(n_{-i,w_i|k}^{(D)} + \beta^{(D)})}{(n_{-i,*|k}^{(D)} + V\beta^{(D)})}. \tag{6}$$

where $z_{-i}^{(D)}$ is the topic assignments for all words, $n_{-i,k|d}$ is the number of words assigned to document topic k in document d, $n_{-i,*|d}$ is the total number of words in document d, $n_{-i,w_i|k}^{(D)}$ is the number of word w_i in document topic k, and $n_{-i,*|k}^{(D)}$ is the total number of words in document topic k. All of them same exclude current assignment of $z_i^{(D)}$.

After a sufficient number of iterations, we can estimate the topic attribution and word distribution as follows:

$$\theta_{ak}^{(A)} = \frac{n_{k|a} + \alpha^{(A)}}{n_{*|a} + K_A\alpha^{(A)}}, \tag{7}$$

$$\theta_{dk}^{(D)} = \frac{n_{k|d} + \alpha^{(D)}}{n_{*|d} + K_D\alpha^{(D)}}, \tag{8}$$

$$\phi_{kw}^{(A)} = \frac{n_{w|k}^{(A)} + \beta^{(A)}}{n_{*|k}^{(A)} + V\beta^{(A)}}, \tag{9}$$

$$\phi_{kw}^{(D)} = \frac{n_{w|k}^{(D)} + \beta^{(D)}}{n_{*|k}^{(D)} + V\beta^{(D)}}. \tag{10}$$

According to Eqs. (3), (4), the process that words are generated from our model is shown in Algorithm 1. Then we estimate author topic distribution and document topic distribution by Eqs. (7), (8) and estimate the topic word distribution for authors and documents by Eqs. (9), (10).
And the expected values for the corpus author distribution are:

$$\chi_a = \frac{1 + d_a}{A + D}. \tag{11}$$

where d_a is the number of documents belonging to the author a.

3.3 Authorship Attribution by Topic-Based Similarity

In the classification phase, we assume all test documents are written by a same unknown author, so no sampling would be required to obtain author topic distribution. Besides, we consider the word distribution of each document topic to be observed in the training phase and use Eq. (10) for getting its expected value. Then we carry on collapsed Gibbs sampling by the following conditional distribution for a given test document $\widetilde{d}$:

$$p(z_i^{(D)}|z_{-i}, \widetilde{\mathbf{D}}, \widetilde{d}) \propto \frac{(n_{-i,k|\widetilde{d}} + \alpha^{(D)})}{(n_{-i,*|\widetilde{d}} + K_D\alpha^{(D)})} \phi^{(D)}_{k\widetilde{w}_i}. \tag{12}$$

where $n_{-i,k|\widetilde{d}}$ is the number of words assigned to document topic k in document $\widetilde{d}$, and $n_{-i,*|\widetilde{d}}$ is the number of words in $\widetilde{d}$. All of them exclude current assignment of $z_i^{(D)}$. Finally, we will get test document's topic distribution by Eq. (8).

Algorithm 1. Topic assignment for words

Input:
K_A: the number of author topic;
K_D: the number of document topic;
$\alpha^{(A)}$: a single-valued hyper-parameter for $\theta^{(A)}$;
$\alpha^{(D)}$: a single-valued hyper-parameter for $\theta^{(D)}$;
$\beta^{(A)}$: a single-valued hyper-parameter for $\phi^{(A)}$;
$\beta^{(D)}$: a single-valued hyper-parameter for $\phi^{(D)}$;
D: the training documents matrix;
Output: multinomial parameters $\theta^{(A)}$, $\theta^{(D)}$, $\phi^{(A)}$ and $\phi^{(D)}$;
procedure Topic Assignment For Words
    Randomly initialize the topic assignments for all words;
    while not finished **do**
        for all documents $d \in [1,D]$ **do**
            for all words $w_{di} \in [1,N_d]$ in document d **do**
                Draw author topic $z_{di}^{(A)}$ by equation (5);
                Update $n_{k|a}$, $n_{w|k}^{(A)}$ and $n_{*|k}^{(A)}$;
                Draw document topic $z_{di}^{(D)}$ by equation (6);
                Update $n_{k|d}$, $n_{w|k}^{(D)}$ and $n_{*|k}^{(D)}$;
            end for
        end for
        if converged and L sampling iterations since last read out **then**
            read out parameter set $\theta^{(A)}$ according to Equation (7);
            read out parameter set $\theta^{(D)}$ according to Equation (8);
            read out parameter set $\phi^{(A)}$ according to Equation (9);
            read out parameter set $\phi^{(D)}$ according to Equation (10);
        end if
    end while
end procedure

To find the author of test documents, we calculate the similarity between the test document $\widetilde{d}$ and the training document d based on the topic probability distribution as follows:

$$Similarity(\widetilde{d}, d) = \chi_a \prod_{i=1}^{N_{\widetilde{d}}} \sum_{k=1}^{K_A} \theta_{ak}^{(A)} \phi_{k\widetilde{w}_i}^{(A)} - \sqrt{\sum_{k=1}^{K_D} (\widetilde{\theta}_{\widetilde{d}k}^{(D)} - \theta_{dk}^{(D)})^2}. \quad (13)$$

In the Eq. (13), we first calculate the probability of $\widetilde{d}$ written by the author of d. Then we calculate the Euclidean distance between $\widetilde{d}$ and d based on document topic distribution. The difference between them is the final similarity between $\widetilde{d}$ and d. After calculating all similarity, we assign $\widetilde{d}$ to the author of document d with the largest similarity value. In this way, both authors and documents play important roles in authorship attribution, which is never considered before to our knowledge. The classification process is shown in Algorithm 2.

Algorithm 2. Author assignment for test documents

Input:
A_d:the author of training documents set
D: the training documents matrix;
$\widetilde{\mathbf{D}}$: the test documents matrix;
$\theta^{(A)}$: the author topic distribution;
$\theta^{(D)}$: the document topic distribution in the training documents **D**;
$\theta^{(\widetilde{D})}$: the document topic distribution in the test documents $\widetilde{\mathbf{D}}$;
$\phi^{(A)}$: the author topic word distribution;
D: the corpus matrix;
Output: *result*: authors of test documents $\widetilde{\mathbf{D}}$;

```
procedure AUTHOR ASSIGNMENT FOR TEST DOCUMENTS
    for all test documents d̃_i ∈ [1,D̃] do
        for all training documents d_j ∈ [1,D] do
            similarity_{i,j}=Similarity(d̃_i,d_j) by equation (13)
            if similarity_{i,j} > currentMaxSimilarity then
                currentMaxSimilarity = similarity_{i,j};
                currentAuthor = a_{d_j};
            end if
        end for
        a_{d̃_i}=currentAuthor;
    end for
end procedure
```

3.4 ADT vs DADT

ADT seems a little similar to DADT because both of them build the topic model for the corpus at the document level and the author level, but there are two key differences between them.

Firstly, we treat every word of the corpus equally at both two levels in ADT, which is more suitable for short texts. When we apply DADT to short texts, the words of the corpus are divided into two disjoint sets. In other words, each word of the corpus has an author topic and a document topic in our model, but it only has an author topic or a document topic in DADT.

Secondly, considering the effect of both authors and documents, we propose a new classification algorithm to find the most probable author of the anonymous texts. In this way, both authors and documents play important roles in authorship attribution for short texts.

Table 1. Detail properties of the datasets.

Dataset	Pan'11	Blog
Authors	71	136
Avg.texts	143	61
Avg.words $\pm$ stedv	39 ± 33	57 ± 52

4 Experiments

4.1 Datasets

We use two public datasets for the experiment evaluations. The first one is **Pan'11 emails** [8]: 11936 emails with 72 authors. The second one is **Blog** [9]: 678161 blogs with 19320 authors. We want to figure out the problem of authorship attribution for short texts. Hence we delete emails and blogs which have more than 1000 characters. Besides, considering the time and space constraints, we only select 136 prolific authors from blogs. After pre-processing, some statistics about the datasets are provided in Table 1.

A quick analysis of the two datasets shows that the blog dataset is noisier than the Pan'11 dataset. Compared to Pan'11, Blog has more authors but fewer documents. The average length of documents from the blog is also longer than the Pan'11 dataset. In addition, we remove one author's documents from Pan'11 because they all have more than 1000 characters. We use both datasets to evaluate our ADT model's stability of performance on different scales of the corpus. We indicate the significance using t-test, two-tailed, p-value <0.05.

4.2 Baselines

Character 4-gram is an effective feature for authorship attribution, but we find it only has a positive effect on SCAP [10], and FS [3]. In topic models, like AT, DADT [14] and our proposed method ADT, or SVM, compared with word feature, character 4-gram does not perform better but needs more time to train the topic models. Considering the cost of time, we model the corpus with character

4-gram features in SCAP, and FS, and we model the corpus with word features in topic models and SVM.

SCAP. We build the author profiles for each of the candidates from the corpus and calculate the Jaccard similarity between a given text and the profile texts of authors to find the most similar author.

FS. We test different values for the parameter of the feature sampling method k. We find when we randomly sample 40% features from the feature set and k = 100, the best performance is achieved.

SVM. SVM is widely used in text classification. We use linear SVM in a one-versus-all setup, as implemented in Weka [17].

AT. We calculate the probability of each test text for each author by the inferred Author-Topic model and return the most probable author. Compared to LDA-H [7], AT significantly performs better when the corpus has tens of authors [14].

DADT. DADT is our most important baseline for comparison. When we get an inferred DADT model, the model assumes that test texts are written by a new author, and use the given model to infer the author/document topic ratio and the document topic distribution. Then we calculate the probability of each author for test texts and return the most probable author.

Table 2. Accuracy with different topic values on ADT and DADT. The best performance is highlighted in bold.

K_A/K_D	40/10	90/10	140/10	190/10	240/10
Pan'11					
ADT	49.2%	54.0%	54.2%	54.2%	**54.7%**
DADT	45.2%	51.2%	51.2%	51.1%	**51.3%**
Blog					
ADT	36.9%	45.6%	48.5%	**49.2%**	48.9%
DADT	32.9%	41.1%	44.4%	**45.7%**	45.5%

4.3 Experiment Setup

In all experiments, ten-fold cross-validation is carried out on both two datasets. We first trained all the methods on the training set, tuned the parameter according to the results on the test set to get the best performance. We use classification accuracy which is the percentage of test texts that are assigned to the correct author for evaluating results.

To train the topic models, we use collapsed Gibbs sampling with a burn-in of 1000 iterations. In all of the experiments with topic models, we retain 100 of samples with the spacing of 1 iterations. According to Eqs. (7), (8), (9) and (10), we estimate the author topic distribution, document topic distribution,

author word distribution and document word distribution from training samples for our models. Then we average them to get stable parameter estimates for the models. In classification phase, we use a burn-in of 100 iterations and average the parameter estimates over the next 100 iterations to get the final document topic distribution for test texts. In addition, we set $\alpha^{(A)}$ and $\alpha^{(D)}$ to $min\{0.1, 5/K_A\}$ and $min\{0.1, 5/K_D\}$, $\beta^{(A)}$ and $\beta^{(D)}$ to 0.01 for getting the best performance.

4.4 Results

Table 2 shows the accuracy of ADT and DADT with different values of author topic K_A on Pan'11 and Blog. We find that when we increase the value of document topic K_D, it does not improve the performance of both models, so we set K_D to 10 in all experiments. Compared to Pan'11, Blog needs more author topics to obtain the relatively stable ability of classification for the models. The results indicate that the performance of ADT outperforms DADT with all different values of K_A.

Table 3. Accuracy with different models on two datasets. The best method is highlighted in bold.

Dataset	Pan'11	Blog
ADT	**54.7%**	**49.2%**
DADT	51.3%	45.7%
AT	47.8%	43.6%
SVM	47.0%	30.8%
FS	43.9%	43.4%
SCAP	22.9%	26.5%

The results of our model and other baselines on Pan'11 and Blog are shown in Table 3. Except for SCAP, all the methods yield relatively low accuracies on Blog, but the accuracy of FS only decreases a little on Blog. As we can see, three kinds of topic models achieve better performance on both two datasets than other methods. This is because topic model can handle the problem of data sparsity on the short texts. Compared to the state of the art, our ADT model obtains a 6.63% improvement on Pan'11 and a 7.66% improvement on Blog, which demonstrates the effectiveness and stability of ADT.

5 Conclusions

The main goal of this paper is to deal with the problem of authorship attribution for short texts. We propose a novel model ADT which combines Author-Topic model and traditional LDA to build the model for the corpus due to the irregular

usage and data sparsity of short texts. Compared to the traditional methods, ADT tries to make use of the effect of training texts on authorship attribution, and we propose a new classification algorithm which combines the influence of both authors and texts to calculate the similarity between texts for finding the most probable author of test texts instead of only aggregating texts into a lengthy pseudo-document. On both real-world datasets, our proposed method performs significantly better than the current state of the art. In the future, we would like to combine the time factor and topic model to improve the performance of authorship attribution for short texts.

Acknowledgements. This work is supported by the National Natural Science Foundation of China [grant number 61772289] and the Fundamental Research Funds for the Central Universities.

References

1. Stamatatos, E.: A survey of modern authorship attribution methods. J. Assoc. Inf. Sci. Technol. **60**(3), 538–556 (2009)
2. Abbasi, A., Chen, H.: Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. **26**(2), 1–29 (2008)
3. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Lang. Resour. Eval. **45**(1), 83–94 (2011)
4. Azarbonyad, H., Dehghani, M., Marx, M., Kamps, J.: Time-aware authorship attribution for short text streams. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 727–730 (2015)
5. Hofmann, T.: Probabilistic latent semantic indexing. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
7. Seroussi, Y., Zukerman, I., Bohnert, F.: Authorship attribution with latent Dirichlet allocation. In: Fifteenth Conference on Computational Natural Language Learning, pp. 181–189 (2011)
8. Argamon, S., Juola, P.: Overview of the international authorship identification competition at PAN-2011. In: Petras, V., Forner, P., Clough, P. (eds.) Notebook Papers of CLEF 2011 Labs and Workshops, Amsterdam, Netherlands, 19–22 September 2011
9. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. Front. Inf. Technol. Electron. Eng. **274**(s 1–2), 199–205 (2006)
10. Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C.E., Howald, B.S.: Identifying authorship by byte-level N-grams: the source code author profile (SCAP) method. Int. J. Digit. Evid. **6**(1), 1–18 (2007)
11. Koppel, M., Schler, J., Argamon, S., Messeri, E.: Authorship attribution with thousands of candidate authors. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 659–660 (2006)

12. Sousa Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., Maia, B.: 'twazn me!!!;(' Automatic authorship analysis of micro-blogging messages. In: Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 161–168. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22327-3_16
13. Schwartz, R., Tsur, O., Rappoport, A., Koppel, M.: Authorship attribution of micro-messages. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1880–1891 (2013)
14. Seroussi, Y., Bohnert, F., Zukerman, I.: Authorship attribution with author-aware topic models. In: Meeting of the Association for Computational Linguistics: Short Papers, pp. 264–269 (2012)
15. Yang, M., Zhu, D., Tang, Y., Wang, J.: Authorship attribution with topic drift model. In: AAAI, pp. 5015–5016 (2017)
16. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. **101**(Suppl 1), 5228 (2004)
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)