Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

A simple but Powerful E-mail Authorship Attribution System

A. K. M. Mustafizur Rahman Khan[+]

[1] Division of Applied Mathematics and Computer Science
King Abdullah University of Science & Technology
23955-6900, Thuwal, Jeddah, K.S.A

Abstract. There has been an alarming increase of cybercrimes recently. In many cases, e-mails are used by criminals as favorite weapons. Therefore e-mail authorship attribution system is a necessity to prevent such crimes and identify criminals. Many successful techniques have been invented for authorship attribution of literary works. Due to informality and short size of e-mails, these techniques cannot perform well in e-mail authorship attribution. A few recent research works have proposed very complex methods which use stylometric features and grammatical rules etc. We propose a simple but powerful and robust method based on ensemble method and Naïve Bayes classifier. Our method is grammar independent thus can deal with e-mails written in almost all languages. In order to evaluate our method, we used Enron corpus which is a collection of real world e-mails. Although simple, our method outperforms the existing methods. Authorship attribution accuracy heavily depends on length of the concerned e-mail. But none of the previous works addressed this issue. We did a comprehensive study of relation between accuracy of e-mail authorship attribution and e-mail length.

Keywords: E-mail authorship attribution, Enron corpus, text analysis, Naïve Bayes.

1. Introduction

With the spread of World Wide Web, E-mail has changed our way of written communication. The increase in e-mail traffic comes also with an increase in the use of e-mails for illegitimate purposes such as phishing, spamming, e-mail bombing, threatening, cyber bullying, racial vilification, child pornography, and sexual harassment etc. Like other criminals, the cyber criminals attempt to hide their true identity. For example, in phishing, a person may try to impersonate a manager or a financial adviser to obtain clients' secret information. Presently, there is no adequate proactive mechanism to prevent e-mail misuses. In this situation, e-mail authorship attribution can help e-mail service provider to detect a hack and the recipient to detect a fraud.

Authorship analysis has been very successful for literary and conventional writings. Specially, stylometric features have been extensively used for long time [1]. More than 1000 stylometric features comprising of lexical, syntactic, structural, content-specific, and idiosyncratic characteristics have been used in various studies [1, 2, 3, 4, 5, 6]. This trend of using stylometric features is also found in e-mail authorship attribution studies. F. Iqbal et al. used 292 stylometric features [8]. B. Allison et al. generated the grammar rules used in the e-mail and used these as features [7]. Literary documents usually have a definite syntactic and semantic structure and are usually large in size. In contrast, e-mails are short in size, more interactive and informal in style and usually do not follow definite syntactic or grammatical rules. As a result, value of these stylometric and grammatical features cannot be calculated reliably. Moreover, due to unreliable values, these large numbers of features remain redundant and lower accuracy of an authorship attribution system by creating noise. Therefore, techniques which are very successful in literary and traditional works are not suitable in e-mail authorship attribution.

[+] Corresponding author. Tel.: + 8801911765187.
E-mail address: mustafiz_46@yahoo.com.

In our study, we have generated bag of words from training e-mails for each author and a bag of words for background. Since we think that longer length n-grams have more potential than single words, we have generated bag of bigrams of words for each author and background. We have implemented four Naïve Bayes classifiers and predicted the author of the test e-mail by their votes.

2. Our Method

We assume that we have a collection of e-mails of each suspected author. These e-mails are the training e-mails. We generate a bag of words from training e-mails for each author. Consecutive sequences of English alphabetic characters are treated as a word. We remove all the words which appear only once in the training e-mails of an author considering them as noise. Similarly, we generate a bag of bigrams for each author and remove noise. Then we generate a bag of words and a bag of bigrams from all training e-mails for background. Again we remove the words and bigrams which appear only once. We do not use longer n-grams because n-grams frequency decreases exponentially with n. Therefore, very few longer n-grams from a short test e-mail can be found in training e-mails resulting in unreliable statistics. We construct four Naïve Bayes classifiers based on bag of words, bag of bigrams, bag of words considering background and bag of bigrams considering background. When a test e-mail is presented we generate a bag of words and a bag of biagrams from it. But in this case we keep all words and bigrams. We calculate the likelihood (l_{1i}) of a test e-mail with i-th author's training e-mails considering single words by the following algorithm. Likelihood (l_{2i}) is calculated considering bigrams instead of words by similar algorithm.

Algorithm 1:

INPUT: A test e-mail (E), bag of words of i-th author (B_i)

OUTPUT: likelihood(l_{1i})

1: l_{1i}**=0**
2: **foreach** word(w_k) in E
3: **if** (w_k is present in B_i)
4: count= count of w_k in B_i
5: **else**
6: count=default count
7: **end if**
8: l_{1i}=l_{1i}+log(count/size of B_i)
9: **end foreach**
10: **return** l_{1i}

Then we calculate likelihood(l_{3i}) considering single words and background by algorithm 2.

Algorithm 2:

INPUT: A test e-mail (E), bag of words of i-th author (B_i)

INPUT: Bag of words of background (B_{bk})

OUTPUT: likelihood(l_{3i})

1: l_{3i}=0
2: **foreach** word(w_k) in E
3: **if** (w_k is present in B_{bk})
4: count_bk = (1+ count of w_k in B_{bk})/ size of B_{bk}
5: **if** (w_k is present in B_i)
6: count= (1+count of w_k in B_i)/ size of B_i
7: **else**
8: count= 1/ size of B_{bk}
9: **end if**

10: **end if**

11: l_{3i}=l_{3i}+log(count/count_bk)

12: **end foreach**

13: **return** l_{3i}

We calculate likelihood (l_{4i}) considering bigrams and background by similar algorithm. There may be some words or diagrams in test e-mail which are not present in a particular author's bag of words/bigrams but present in the background bag. This particular author avoids these words. We can infer that these words are written by another author. We used background bag of words/bigrams in order to use this logic.

We create 4 sets of likelihood values by calculating four likelihood values for each author. These 4 sets of likelihood values can be viewed as output of 4 Naïve Bayes classifiers. Then we perform z-score normalization in each set. Next we add these 4 normalized likelihood values of each author. We attribute authorship to the author who corresponds to maximum sum of normalized likelihood values.

3. Experimental Evaluation

We performed experiments on the Enron e-mail corpus, made available by MIT, in order to evaluate our method. The dataset contains about half a million real-life e-mails from 158 employees of Enron Corporation. We randomly selected 10 employees who sent a substantial amount of e-mails. The selected employees are Phillip K Allen, John Arnold, Lynn Blair, Sally Beck, Michelle Cash, David W Delainey, Chris Dorland, Vince J Kaminski, Richard B Sanders and Bill Williams III. We scanned their "_sent_items", "sent", and "sent_mail" folders and collected all e-mails. Although the dataset is organized in folders, in many cases several copies of an e-mail were found in different folders. Duplicate e-mails were identified by comparing time stamps of the e-mails. We kept only one of the copies. The e-mails contain a lot of noise. Many e-mails are just forwarded. These forwarded e-mails were also removed. However, the texts which were quoted from others or part of advertisement were not removed. No effort was done to correct spelling.

We used 10 fold cross validation approach. We reserved 90% of the e-mails for training and 10% for testing. Like B. Allison et al.[7] we discarded the e-mails which are shorter than 20 words. The mode length of the test e-mails was 23 and median length was 46. In total 6109 e-mails were used for testing.

In order to evaluate effect of e-mail length we did another experiment. In this case also we used 10 fold cross validation. We extracted texts of all test e-mails and dumped them in one file. Then we slide a window of a fixed length over the whole file one word by one word and selected the text within the window. We attributed authorship to the selected text. We tried with different window lengths and calculated accuracy.

4. Results and Discussions

Tab.1 shows us the result of 10-way authorship attribution experiment. Accuracy of our method is 86.92% which is much better then accuracy achieved by F. Iqbal et al.[4]. They reported 77% accuracy in a 10-way authorship attribution experiment. Note that they selected 200 e-mails written by 10 authors. They neither reported selected authors' names nor mentioned e-mail selection criterion.

Table. 1: Error count of 10-way authorship attribution experiment.

Number of	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Total
Test e-mail	579	742	509	576	583	530	562	552	646	830	6109
Error	80	97	71	75	60	72	67	67	91	119	799

In order to compare with B. Allison et al.[7] we removed Bill Williams III from author list and conducted a 9-way authorship attribution experiment. Tab.2 shows the results. In this experiment accuracy of our method is 87.50% which exceeds the best accuracy reported in B. Allison et al. They compared several methods and reported 87.05% as the best accuracy. Note that they did not report selected authors' names.

Table. 2: Error count of 9-way authorship attribution experiment.

Number of	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Total
Test e-mail	548	710	480	531	552	497	529	541	610	801	5799
Error	70	87	62	63	53	66	63	62	86	113	725

Table 3 shows us error rate at different window lengths from our second experiment described in previous section. This experiment was done with 10 authors.

Table. 3: Error Window size Vs error rate.

Window size	25	50	75	100	125	150	175	200
Error rate	0.2239	0.1155	0.0706	0.0479	0.0332	0.0246	0.0183	0.0132

Our method achieves very high accuracy when the window length is higher than 100. Fig.1 shows us the relation between window length and error rate. From this result we can conclude that, possibility of error in an e-mail's authorship attribution decreases logarithmically with the e-mails length.

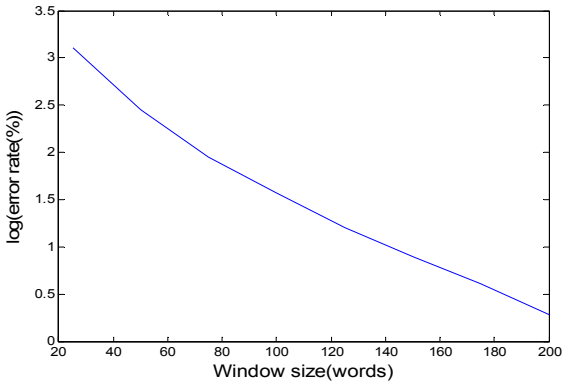

Fig.1 Window size vs error rate.

5. Summaries

In this paper, we have proposed a novel method of e-mail authorship attribution. Our method outperforms existing methods. It achieves very high accuracy when the e-mail is longer than 100 words. Since our method does not use any stylometric and grammatical feature, it is applicable to e-mails written in other languages. Our method is very easy to implement because it does not require any complex feature extraction.

6. References

[1] T.C. Mendenhall. The characteristic curves of composition. *Science*, 11(11):237–249, 1887.

[2] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2), 2008.

[3] M. Corney, O. Vel, A. Anderson, and G. Mohay. Gender-preferential text mining of e-mail discourse. In proc. *18th Annual Computer Security Applications Conference.* 2002, PP: 21–27.

[4] F. Iqbal, R. Hadjidj, B. C. M. Fung, and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*, 5:42–51, 2008.

[5] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology,* 57(3):378–393, February 2006.

[6] R. Zheng, Y. Qin, Z. Huang, and H. Chen. Authorship analysis in cybercrime investigation. In proc. *1st NSF/NIJ Symposium, ISI Springer-Verlag.,* PP: 59–73, 2003.

[7] B. Allison and L. Guthrie, " Authorship attribution of e-mail: comparing classifiers over a new corpus of evaluation," in Proceedings of the *Sixth International Language Resources and Evaluation (LREC'08)*, May 28-30, 2008, Marrakech, Morocco.

[8] F. Iqbal, L.A. Khan, B.C.M. Fung, M. Debbabi: E-mail authorship verification for forensic investigation. In: Proc. of the *2010 ACM Symposium on Applied Computing.* pp. 1591–1598. SAC '10, ACM, New York, NY, USA (2010).