EPSMS and the Document Occurrence Representation for Authorship Identification[*]

Notebook for PAN at CLEF 2011

Hugo Jair Escalante

Graduate Program in Systems Engineering,
Universidad Autónoma de Nuevo León,
San Nicolás de los Garza, NL 66450, México
hugo.jair@gmail.com
http://hugojair.org

Abstract This paper describes the participation of the PISIS team in the authorship identification track of PAN'11. We adopted two different strategies for the tasks of authorship attribution and authorship verification. For authorship attribution we performed experiments with a document occurrence representation using a standard classification-based approach. Results obtained with this approach were mixed: in the small data sets distributional representations resulted very helpful, although in the large data sets a simple bag-of-words approach outperformed the document occurrence approach. For authorship verification we adopted a classification-based approach and proposed a modification to Ensemble Particle Swarm Model Selection (EPSMS) for selecting classification models for each task. This approach obtained acceptable performance in two out of the three data sets.

1 Introduction

Authorship attribution (AA) and authorship verification (AV) are two closely related problems that aim at uncovering the writing style of authors [27]. Applications of AA and AV include spam filtering [30], fraud detection [14], computer forensics [18], cyber bullying [23] and plagiarism detection [25]. Because of its wide applicability, mainly in security aspects, the development of automated AA techniques has received much attention recently [16,25,27].

AA is defined as the task of identifying whom, from a set of candidates, is the author of a given document [27]. While AV is the task of deciding whether given text documents were or were not written by a certain author [17]. Effective methods have been proposed for both tasks so far, see for example the methods evaluated and/or reviewed in [28,15,16,25,27]. One of the most popular formulations for AA and AV is that based in supervised machine learning methods, where both problems are faced as classification tasks. More specifically, AA can be faced as one of multiclass classification, with

[*] The author thanks the efforts of the organizers of PAN'11 for providing a valuable forum for the evaluation of authorship identification methods. Also, the author thanks the helpful comments of reviewers.

as many labels as candidate authors [10,13]. AV, on the other hand, can be faced as a binary classification problem [9,14].

This paper describes the approach adopted by the PISIS[1] team for the Authorship Identification track for the PAN[2] Lab at CLEF 2011, see [25] for more information on the PAN competition and workshop series. We adopted classification-based methods for facing both AA and AV tasks. For AA we used standard classification algorithms with a distributional term representation for documents. Intuitively, we want to model the writing style of authors in terms of their association with other documents, as modeled with the document occurrence representation. Experimental results in the PAN'11 Authorship Attribution track show that proposed approach resulted very helpful for the small data set. Results in the large data set are very competitive as well, although we found that a simple bag-of-words representation and a nonlinear classifier outperformed the distributional representations.

For AV we used a method called Ensemble Particle Swarm Model Selection [6] for building ad-hoc classifiers for each AV task. We used sample documents by the author as positive examples and documents written by other authors as negative examples. In order to obtain stable predictions we adopted a meta-ensemble approach that combines the outputs of several runs of the model selection technique. Documents are ranked by probability that they were written by the author of interest. Experimental results in the PAN'11 Authorship Attribution track show that the proposed approach resulted very effective for 2 out of 3 data sets, although there are several aspects of the proposed methodology that still can be improved.

The rest of this working note describes in detail the methodologies adopted for the AA and AV tasks, reports the results obtained with them and summarizes our main findings. Before describing the proposed methodology we briefly review related work on AA and AV in the next section.

2 Related work

In the classification-based approach to AA and AV sample documents written by each author are considered instances of an usual classification problem [16,27]. Learning algorithms that have been used for AA and AV include support vector machine (SVM) [10,17,13] and variants thereon [24], neural networks [29], Bayesian classifiers [2], decision tree methods [16] and similarity based techniques [18,16] among several others.

In the above works the same learning algorithm have been used for building the classification models of all of the authors in consideration. An exception is the work by Escalante et al. [9], where particle swarm model selection (PSMS) was used for building specific classification models for each author. The hypothesis of that work is that by considering specific methods for preprocessing, feature selection and classification for each author will increase the classification performance. Satisfactory performance was obtained in the task of AV (i.e., binary classification), although AA performance (i.e., multiclass classification) was limited (because of the incompatibility of scales for the outputs of different models, see [8]). Since PSMS has proved to be very effective for

[1] http://pisis.fime.uanl.mx/

[2] http://www.webis.de/research/workshopseries/pan-11/

diverse binary classification tasks [5,6,9] in this paper we adopt a modified PSMS for the AV task and we used standard learning algorithms for the AA task in the PAN'11 Authorship Identification track.

While standard learning algorithms have been used for AA and AV, a wide diversity of features have been used for representing documents, including, character, lexical, syntactical, grammatical and semantic, among others [12,16,27]. Nevertheless, the most used representation is still the one based on the bag-of-words formulation. In particular, the bag-of-words formulation using character n-grams has terms have been successfully used by several researchers [9,10,13,22]. In this paper we adopted an extended bag-of-words representation for documents called the document occurrence representation (DOR) [19]. Under DOR documents are represented by a distribution of occurrences over other documents in the corpus, in such a way that documents are represented by their context. DOR has been successfully used in term clustering [21], word sense disambiguation [11] and multimedia image retrieval [7].

3 Authorship verification

Three AV tasks we evaluated in the PAN'11 authorship identification track. For each task organizers provide sample documents written by the author (training set) and documents written by the author and other authors (validation set). The developed method was tested in documents from the test set (of course, labels in the test set were not available to participants during the competition). Since both, training and validation data are available during development we merged the documents in the training and validation sets for training our method. Table 1 shows the number of documents written and not written by the author of each data set in the training, validation and test sets.

Table 1. Number of documents written (**Y**) and not written by the author (**N**) in the training, validation and test sets.

Data set	**Training**	**Validation**		**Test**	
	Y	**Y**	**N**	**Y**	**N**
Verify-1+	42	3	104	3	92
Verify-2+	55	3	95	5	101
Verify-3+	47	3	100	4	89

3.1 Features

In our approach to AV we used documents in the training and validation sets as training data for training a classifier that discriminates between documents written by the author and documents written by any other author. Documents were represented by their bag-of-words using character n-grams as terms, with $n = 3$. Spaces and punctuation marks were considered characters. We did not use the distributional term representation for this task because of the small number of documents in the training and validation sets, see Section 4.

3.2 Classification approach

Once that documents are represented by their bag-of-words we used Ensemble Particle Swarm Model Selection (EPSMS) for the selection of classification models for each data set. EPSMS is a method for the automatic selection of binary classification models [6]. In a nutshell, EPSMS searches for the best ensemble method that can be generated by using the methods available in a machine learning toolbox [3]. An ensemble is a classification model that combines the outputs of several classifiers. Under certain conditions, it has been shown that ensembles can achieve better performance than individual models [3]. In previous work we have shown that EPSMS can select very effective ensemble classification models [6,4]. A distinctive feature of EPSMS is that each of the members of the ensemble is a method that differs in terms of preprocessing method, feature selection technique and learning algorithm. The heterogeneity of the considered models (diversity) together with the competitive accuracy (performance) of models guarantee selecting very effective classification models. See [6,4,5] for further details on EPSMS.

For each AV data set we provide as input to EPSMS the training+validation data and EPSMS returns a ensemble classifier. Although EPSMS provides very stable classification models [6,4] in this work we wanted to obtain even more stable models. Therefore, we adopted a meta-ensemble approach in which the outputs of several ensembles (each one selected with EPSMS) were combined. The intuition behind this technique is that by running EPSMS several times and combining the outputs of the corresponding methods we could more stable predictions. Stability is very important in EPSMS as this method is based in a heuristic search method, besides the search space contains many local minima.

The meta-ensemble approach is as follows. For each AV data set we ran EPSMS 5 times. Then the selected ensembles were applied to the test data set. As a result we have for each test document the five outputs provided by the 5 ensembles. The output of each ensemble is a real number between $[0, 1]$ expressing the confidence that the sample belongs to the positive class. The outputs of each ensemble are sorted in descending order in such a way that test documents that are more likely to belong to the positive class (i.e., documents written by the author) are ranked in the first positions. For each ranking we keep the top-10 ranked documents. Then, for each document in the union of the 50 documents we count the number of rankings in which they appeared within the top-10 positions(a number between 1 and 5). Finally, we sort the test documents by this number and assign the positive label to the top 10 ranked documents.

Our hypothesis with this method is that if the document is likely to be written by the author it is very likely that the document will receive a high score from several EPSMS ensembles. Documents not written by the author of interest may appear in the top ranked documents for one or two ensembles, although it is reasonable to assume that the top ranked documents are those with more chances to be written by the author.

[3] `http://clopinet.com/CLOP`

The choice of the top 10 ranked documents was done by analyzing the outputs of the different ensembles. We found that after 10 documents most of the documents received very similar scores in the test set.

4 Authorship attribution

As mentioned in the related work section, the performance of PSMS [5] and EPSMS [8,4] for multiclass classification models is not as good as for binary classification tasks. Therefore, we decided to adopt a different approach for the AA task. In particular, we focused on the evaluation of an extended bag-of-words representation for documents and used a standard classification model. Table 2 summarizes the main statistics of the AA data sets for the PAN'11 Authorship Identification track.

Table 2. Description of the AA data sets. Standard data sets are those not including additional authors not available during training. While plus data sets $(+)$ may include documents written by authors that were not represented in the training set. For each data set and each partition we show the number of documents and between parentheses the number of classes.

Data set	Training	Validation	Test
Small	3001 (26)	518 (23)	495 (23)
Small+	3001 (26)	601 (43)	634 (45)
Large	9337 (72)	1298 (66)	1300 (64)
Large+	9337 (72)	144 (86)	1416 (87)

4.1 Features

The bag-of-words representation using character n-grams as terms is among the most used representations for documents in AA [9,10,13,16,22,27]. Despite the fact that acceptable performance has been obtained with such representation in AA, we think that results obtained with such representation can be improved by adopting extended representations. Several extensions to the bag-of-words approach has been proposed in closely related fields as information retrieval [1], computational linguistics [19] and machine learning [20]. In this work we explore the suitability of the document occurrence representation (DOR) for document representation in AA.

DOR is a distributional term representation in which a document is represented by a distribution of occurrences over other documents in the same corpus [19]. Intuitively, a document is represented by its context. The process for obtaining the DOR representation for documents is as follows. First, each term in the vocabulary is first represented as a distribution of occurrences over documents. Next, each document is then represented by a combination of the representations of terms that occur in the document.

DOR is considered the dual of the *tf-idf* representation for representing documents: as documents can be represented by a distribution over the terms, terms can be represented by a distribution over documents. Each term t_j in the vocabulary V is represented by a vector of weights $\mathbf{w}_j^{dor} =< w_{j,1}^{dor}, \ldots, w_{j,N}^{dor} >$, where N is the number of documents in the collection and $0 \leq w_{j,k}^{dor} \leq 1$ represents the contribution of document $\mathbf{d}_k$ to the representation of t_j. Specifically, we consider the following weighting scheme [19]:

$$\mathbf{w}^{dor}(t_j, \mathbf{d}_k) = df(t_j, \mathbf{d}_k) \times \log\big(\frac{|V|}{N_k}\big) \tag{1}$$

where N_k is the number of different terms that appear in document $\mathbf{d}_k$ and $df(t_j, \mathbf{d}_k)$ is given by:

$$df(t_j, \mathbf{d}_k) = \begin{cases} 1 + log\big(\#(t_j, \mathbf{d}_k)\big) \text{ if } \#(t_j, \mathbf{d}_k) > 0 \\ \quad\quad\quad\quad\quad 0 \text{ otherwise} \end{cases} \tag{2}$$

where $\#(t_j, \mathbf{d}_k)$ denotes the number of times term t_j occurs in document $\mathbf{d}_k$. The weights are normalized using cosine normalization. Intuitively, the more frequent the term t_j occurs in document $\mathbf{d}_k$, the more important $\mathbf{d}_k$ is to characterize the semantics of t_j; on the other hand, the more different terms occur in $\mathbf{d}_k$, the less it contributes to characterize the semantics of t_j.

Once that each term is represented according to Formula (1) each document is represented by the unweighed sum of the representations of terms that appear in the document. In this way, a document is represented as a distribution of occurrences over other documents in the collection. Our hypothesis on the use of DOR for AA is that the expanded representations are more descriptive than the usual bag-of-words approach. We did not use this representation for the AV task because the number of documents in the different tasks are very small (see Table 1), which resulted in very low dimensional representations of documents.

4.2 Classification approach

For classification we used the neural network classifier implemented in the CLOP toolbox [26]. We selected this classifier after performing a preliminary evaluation of several classification algorithms. We found that the combination of DOR representation and neural network classifier achieved the highest performance in the validation data sets. For the standard data sets, see Table 2, we used a straight multiclass classifier (one-vs-all approach), where a class corresponds to an author. For the plus data sets (i.e., data sets that contain documents not written by any author in the training set). We used a multiclass classifier with an extra class: unknown author. We just considered documents not written by any author in the training set as another author. Recall, we used training+validation data for training the classifiers.

5 Evaluation

This section reports the results obtained with the proposed methods in the authorship attribution track of PAN'11. We first analyze the performance of the AV methods and then that of the AA techniques.

5.1 Authorship verification

Table 3 shows the results obtained in the AV data sets. The results are mixed: our EPSMS approach obtained the first position in the second data set, although it was ranked ninth in the third data set. For the data set Verify-1, a single document written by the author (out of three available) was identified, this document was ranked second according to the weights generated with the meta-ensemble approach. The other two relevant documents did not appear in the top ranked documents for any of the 5 ensembles. For the Verify-2 data set 4 out of 5 documents were identified by the EPSMS approach, while no author was correctly identified for the Verify-3 data set 3.

Table 3. Experimental results in the AV task. We show Precision, Recall and F1 measure.

Data set	**Precision**	**Recall**	**F1**	**Sum-Ranks**	**Overall Rank**
Verify-1	0.1	0.333	0.154	17	6^{th} out of 10
Verify-2	0.4	0.8	0.533	11	1^{st} out of 10
Verify-3	0	0	0	30	9^{th} out of 10

From Table 1 we can see that the problems are imbalanced and the fact that negative examples (documents written by other authors) are made of documents from different authors further complicated the classification problem. Nevertheless, the results obtained with the proposed formulation are interesting and give evidence that the classification approach to AV can be very effective. We believe the proposed method has potential for this and other binary classification tasks, although we would like to conduct an extensive evaluation of the proposed approach in order to detect what factors influence the performance of the proposed technique. A limitation of the proposed approach is that it ranks documents that are more likely to be written by the author, and then a threshold (top 10-ranked documents) must be used for determining what documents were written by the author. In future work we would like to study alternative formulations for the combination of the outputs from different ensembles.

5.2 Authorship attribution

Table 4 shows the official results obtained by our methods in the AA task. Overall, we can say that results were very competitive. Our entries were above the average performance among other participants. The results were particularly positive in the Small data sets where our method is ranked second and third. Interestingly, the DOR representation resulted more helpful for the data sets that included authors not seen in the training set. Giving evidence that a classification approach for modeling unknown authors can be an effective solution for this AA scenario. The performance in terms of macro and micro average measures was proportional.

In order to evaluate the advantage of the DOR representation over a standard bag-of-words formulation we performed post-competition experiments[4]. We performed exper-

[4] Participants were provided with the labels for test set documents after the competition finished.

Table 4. Experimental results in the AA task. We show Macro Average (**MA**) and Micro Average (**MI**) Precision (**P**), Recall (**R**) and F1 measure (**F1**). Column **Sum-Ranks** shows the sum of ranks across the different measures, we also show the overall ranking achieved by each entry.

Data set	MA-P	MA-R	MA-F1	MI-P	MI-R	MI-F1	Sum-Ranks	Overall Rank
Small	0.676	0.381	0.387	0.709	0.709	0.709	19	3^{rd} out of 17
Small+	0.65	0.201	0.193	0.578	0.573	0.575	16	2^{nd} out of 13
Large	0.608	0.294	0.303	0.508	0.508	0.508	48	8^{th} out of 18
Large+	0.53	0.203	0.191	0.446	0.446	0.446	29	5^{th} out of 13

iments using the same classification-based approach described in Section 4, although using a binary bag-of-words representation with character n-grams as terms. The same neural network with the same (default) parameters as used with the DOR representation. Table 5 shows the performance obtained by the classifier with both representations.

Table 5. Experimental results in the AA task using DOR and the bag-of-words (BOW) representations. We show Macro Average and Micro Average F1 measure, accuracy and the rank that would be obtained by the different representations using only the F1 measure values. For the plus data sets we were unable to reproduce the performance measurements provided by the organizers, therefore we do not show the computed results for those data sets.

Data set	Accuracy		Macro-F1		Micro-F1		Sum-ranks	
	DOR	BOW	DOR	BOW	DOR	BOW	DOR	BOW
Small	70.91	67.88	0.387	0.418	0.709	0.678	3	4
Small+	57.25	55.20	0.709	0.552	0.193	-	2	2
Large	50.76	62.53	0.303	0.463	0.507	0.6254	8	3
Large+	44.56	53.24	0.446	0.532	0.191	-	5	2

Results are mixed: for the Small data sets the DOR representation outperformed the performance of the bag-of-words formulation. While the improvement in terms of accuracy is considerable the ranking of both methods was not significantly affected. On the contrary, in the Large data sets the bag-of-words approach outperformed the DOR representation. The differences in all of the measures are considerable (more than 10% in accuracy). Note that the ranking for the Large data sets are considerably reduced for the bag-of-words approach. This result was somewhat unexpected, as one may think that since in the Large data sets we have more documents available, the DOR representations can be more informative (richer). We think that this results are due to the fact that having more classes there can be an overlap in the representations for documents that belong to different classes. We will try to clarify this behavior in future work. Another issue could be that the number of documents over which compute the DOR representation (and even the selection of which documents are used) can have an important impact into the performance of methods based on this representation.

6 Conclusions

We described the methods adopted for the PAN'11 Authorship Identification track. Different methods were proposed for the attribution (AA) and verification (AV) tasks. For AV we used EPSMS a tool for the automated selection of ensemble classifiers. Our results show that EPSMS is a very competitive method although it still can further improved. In particular we would like to study different ways to determine that a document has/hasnot been written by an author from the outputs of several ensembles selected with EPSMS. For AA we adopted the document occurrence representation and used a standard classifier. We found that in the Small data sets the DOR representation resulted very helpful, although it was not the case for the Large data sets. It is interesting, and somehow disappointing, that a simple bag-of-words representation outperformed the DOR-based approach in the Large data sets. We would like to analyze in more detail the benefits of DOR for AA and what factors affect the performance of methods based on that representation.

References

1. Carrillo, M., Eliasmith, C., Lopez-Lopez, A.: Combining text vector representations for information retrieval. In: Proc. of the 12th International Conference on Text, Speech and Dialogue (TSD). LNCS, vol. 5729, pp. 24–31. Springer (2009)
2. Coyotl-Morales, R.M., Villaseñor-Pineda, L., Montes-y-Gómez, M., Rosso, P.: Authorship attribution using word sequences. In: Proc. of 11th Iberoamerican Congress on Pattern Recognition. LNCS, vol. 4225, pp. 844–852. Springer, Cancun, Mexico (2006)
3. Dietterich, T.: Ensemble methods in machine learning. In: Proc. of the First workshop on Multiple Classifier Systems. LNCS, vol. 1857, pp. 1–15. Springer (2000)
4. Escalante, H.J., Altamirano, L., Gonzalez, J.A., Montes, M., Gomez, P., Reta, C., Reyes, C.A., Rosales, A.: Acute leukemia classification with ensemble particle swarm model selection. Artificial Intelligence in Medicine Submitted (2011)
5. Escalante, H.J., Montes, M., Sucar, E.: Particle swarm model selection. Journal of Machine Learning Research 10(Feb), 405–440 (February 2009)
6. Escalante, H.J., Montes, M., Sucar, E.: Ensemble particle swarm model selection. In: Proc. of the World Congress on Computational Intelligence. pp. 1814–1821. IEEE, Barcelona, Spain (2010)
7. Escalante, H.J., Montes, M., Sucar, E.: Multimodal indexing based on semantic cohesion for image retrieval. Information Retrieval In press (2011)
8. Escalante, H.J., Montes, M., Sucar, L.E.: An energy-based model for image annotation and retrieval. Computer Vision and Image Understanding 115(6), 787–803 (2011)
9. Escalante, H.J., Montes, M., Villaseñor, L.: Particle swarm model selection for authorship verification. In: Proc. of the 14th Iberoamerican Congress on Pattern Recognition. LNCS, vol. 5856, pp. 563–570. Springer, Guadalajara, Mexico (2009)
10. Escalante, H.J., Solorio, T., Montes, M.: Local histograms of character n-grams for authorship attribution. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. pp. 288–298. ACL (2011)
11. Gale, W.A., Church, K.W., Yarowsky, D.: A method for disambiguating word senses in a large corpus. Computers and the Humanities 26(5), 415–439 (1993)
12. Grieve, J.: Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing 22(3), 251–270 (2007)

13. Houvardas, J., Stamatatos, E.: N-gram feature selection for author identification. In: Proc. of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications. LNCS, vol. 4183, pp. 77–86. Springer, Varna, Bulgaria (2006)
14. Iqbal, F., Khan, L.A., Fung, B.C.M., Debbabi, M.: 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), `http://doi.acm.org/10.1145/1774088.1774428`
15. Joula, P.: Authorship attribution. Foundations and Trends in Information Retrieval 1(3), 233Ű334 (2006)
16. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 60, 9–26 (2009)
17. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proc. of the twenty-first international conference on Machine learning. pp. 62–. ICML 04, ACM, New York, NY, USA (2004), `http://doi.acm.org/10.1145/1015330.1015448`
18. Lambers, M., Veenman, C.J.: Forensic authorship attribution using compression distances to prototypes. In: Computational Forensics, Lecture Notes in Computer Science, Volume 5718. ISBN 978-3-642-03520-3. Springer Berlin Heidelberg, 2009, p. 13. LNCS, vol. 5718, pp. 13–24. Springer (2009)
19. Lavelli, A., Sebastiani, F., Zanoli, R.: Distributional term representations: An experimental comparison. In: Proc. of the International Conference of Information and Knowledge Management. pp. 615–624. ACM Press (2005)
20. Lebanon, G., Mao, Y., Dillon, J.: The locally weighted bag of words framework for document representation. Journal of Machine Learning Research 8, 2405–2441 (2007)
21. Lewis, D.D., Croft, W.B.: Term clustering of syntactic phrases. In: Proc. of the 13th International ACM SIGIR Conference on Research and Development in Informaion Retrieval. pp. 385–404. ACM Press, Bruxelles, Belgium (1990)
22. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proc. of the 22nd International Conference on Computational Linguistics. vol. 1, pp. 513–520. ACM Press, Manchester, UK (2008)
23. Pillay, S.R., Solorio, T.: Authorship attribution of web forum posts. In: Proc. of the eCrime Researchers Summit (eCrime), 2010. pp. 1–7. IEEE, Dallas, TX, USA (2010)
24. Plakias, S., Stamatatos, E.: Author identification using a tensor space representation. In: Proc. of the 18th European Conference on Artificial Intelligence. vol. 178, pp. 833–834. IOS Press, Patras, Greece (2008)
25. Potthast, M., Stein, B., Barrón, A., Rosso, P.: An evaluation framework for plagiarism detection. In: Proc. of the 23rd International Conference on Computational Linguistics (COLING 2010). pp. 997–1005. ACL (August 2010)
26. Saffari, A., Guyon, I.: Quickstart guide for CLOP. Tech. rep., Graz University of Technology and Clopinet (May 2006)
27. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
28. Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E.: Proc. of the 3rd international workshop on uncovering plagiarism, authorship, and social software misuse, PAN'09 (2009)
29. Tearle, M., Taylor, K., Demuth, H.: An algorithm for automated authorship attribution using neural networks. Literary and Linguist Computing 23(4), 425–442 (2008)
30. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Multitopic email authorship attribution forensics. In: Proc. of the ACM Conference on Computer Security - Workshop on Data Mining for Security Applications. Philadelphia, PA, USA. (2001)