Author Identification using Latent Dirichlet Allocation*

Hiram Calvo[1,2], Ángel Hernández-Castañeda[1], Jorge García-Flores[2]

[1]Instituto Politécnico Nacional, Center for Computing Research CIC-IPN
Av. J.D. Bátiz e/ M.O. de Mendizábal, 07738, Mexico City, Mexico
{hcalvo, ahernandez}@cic.ipn.mx

[2]Laboratoire d'Informatique de Paris Nord, CNRS, (UMR 7030) Université Paris 13,
Sorbonne Paris Cité, F-93430, Villetaneuse, France
jgflores@lipn.univ-paris13.fr

Abstract. We tackle the task of author identification at PAN 2015 through a Latent Dirichlet Allocation (LDA) model. By using this method, we take into account the vocabulary and context of words at the same time, and after a statistical process find to what extent the relations between words are given in each document; processing a set of documents by LDA returns a set of distributions of topics. Each distribution can be seen as a vector of features and a fingerprint of each document within the collection. We used then a Naïve Bayes classifier on the obtained patterns with different performances. We obtained state-of-the-art performance for English, overtaking the best FS score reported in PAN 2015, while obtaining mixed results for other languages.

1 Introduction

Author verification is an important problem to solve since many tasks require recognizing the author who wrote a specific text. For example, from knowing which author wrote an anonymous book, up to identifying notes of a serial killer. In this paper we deal with an author verification challenge from a more realistic approach. Specifically, the dataset used consists of one to five documents of a known author and one document of an unknown author. The corpus is formed by four subsets in different languages (English, Spanish, Dutch and Greek). The aim is to identify whether a written unknown text was written by the same author who wrote the known texts. It is important to note that this task becomes more difficult when the dataset is composed of short documents; since current approaches are not able to capture effective models with few amounts of words [1]. However, on real cases as forensic field, long texts rarely exist.

* The authors wish to thank the support of the Instituto Politécnico Nacional, (COFAA, SIP) and the Mexican Government (CONACYT, SNI). The first author is currently in a research stay at Laboratoire d'Informatique de Paris Nord, CNRS, Université Paris 13.

Several approaches have been conducted to generate more informative features based on text style. Nevertheless, it is also possible to generate features by extracting lexical, syntactic, semantic information among others. Lexical information is limited to word counts and occurrence of common words. On the other hand, syntactic information is able to obtain, to a certain extent, the context of the words.

In this work we use semantic information to find features that help us to discriminate texts. For this purpose, we create a model by using Latent Dirichlet Allocation (LDA). By using this method, we consider all the vocabulary from all texts at the same time, and, after a statistical process, find to which extent the relations between words are given in each document. LDA is a statistical algorithm which considers a text collection as a topics mixture; then, processing a set of documents by LDA returns a set of topic distributions. Each distribution can be seen as a vector of features and a fingerprint of each document within the collection. We use machine learning algorithms to classify the obtained patterns.

In this work we obtained the following F-measures: 85.5% for English, 76.0% for Spanish, 70.9% for Dutch and 64.0% for Greek.

2 Related work

Several works have attempted the authorship identification challenge by generating different kinds of features [14], [16]. The nature of the dataset can determine the difficulty of the task, i.e., how hard will be to extract appropriate features [19], [20]. In [3] can be seen that, while the number of authors increases and the size of training dataset decreases, classification performance lowers. This sounds logical since, when the size of training data is lower, the identification of helpful features becomes affected.

Many works address author identification through the author's writing style [15], [18]. For instance, in [4], style-based features are compared to the BoW (Bag of Words) method. This study attempts to discriminate authors from texts in the same domain obtained from Twitter. Style markers such as characters, long words, whitespaces, punctuation, hyperlinks, parts of speech, among others, were included. The study findings showed that a style-based approach was more informative than a BoW-based method. However, their best results were obtained when considering two authors, so there was an accuracy decrease when the number of authors was increased. This suggests that, depending on how big is the training set, there will be stylistic features that help to distinguish an author from other, but not from all other authors.

Stylistic features also can be applied to other tasks. In [5], the authors combined features to address two-class problems. This work attempts to obtain style, BoW and syntax features to classify native and non-native English, texts written for conference

or workshop and texts written by male or female. The dataset consists of scientific articles. This kind of texts is more extensive, compared to e-mail, tweets, or other short texts; this could have led to identify non-native written texts with promising accuracy. Nevertheless, long texts not necessarily ensure good results, since classification tasks on venue and gender obtained low accuracy.

The purpose of identifying authorship can vary. For example, Bradley et al. [6] attempt to prove that it is possible to find out which author wrote an unpublished paper (for a conference or journal); they consider only the cited works in them. By using LSA, the authors propose to create a term-document matrix wherein possible authors are considered as documents and authors who are cited are considered as terms. The results of Bradley et al. showed that the blind review system should be examined in greater detail. Another example is the Castro and Lindauer's work [17], with the task of finding out whether Twitter users identity can be uncovered by their writing style. The authors focused in features such as word shape, word length, character frequencies, stop words' frequencies, among others. With an RLSC (Regularized Least Square Classification) algorithm, the authors correctly classified 41% of the tweets.

In the work of Pimas et al. [13], the author verification task is addressed by generating three types of features. The authors extract stylometric, grammatical and statistical features. Our This study is based on PAN 2015 authorship verification challenge. In addition, Pimas et al consider topics distribution as well, but they argue against using it, because the dataset is formed by topic mixtures. A cross validation model (10 folds) shows good performance, but, on the other hand, the model got overfitting using the training and test sets specified in the dataset.

3 Author verification

In this section we present our method for author verification. First, in Section **Error! Reference source not found.** we detail the source of features we use. Next, in Section 3.2 we describe the dataset used in this work for evaluation, and finally in Section 3.3 we give details on our feature vector construction.

3.1 Latent Dirichlet Allocation (LDA)

LDA [8] is a probabilistic generative model for discrete data collections such as texts collection. It represents documents as a mix of different *topics*. Each topic consists of a set of words that keep some link between them. Words, in its turn, can be chosen based on probability. The model assumes that each document is formed word-by-word by randomly selecting a topic and a word for this topic. As a result, each

document can combine different topics. Namely, simplifying things somewhat, the generation process assumed by the LDA consists of the following steps:

1. Determine the number *N* of words in the document according to the Poisson distribution.
2. Choose a mix of topics for the document according to Dirichlet distribution, out of a fix set of *K* topics.
3. Generate each word in the document as follows:
    a) choose a topic;
    b) choose a word in this topic.

Assuming this generative model, LDA analyzes the set of documents to reverse-engineering this process by finding the most likely set of topics of which a document may consist. LDA generates the groups of words (topics) automatically; see Figure 1.

Accordingly, LDA can infer, given a fixed number of topics, how likely is that each topic (set of words) appear in a specific document of a collection. For example, in a collection of documents and 5 latent topics generated with the LDA algorithm, each document would have different distributions of 5 likely topics. That also means that vectors of 5 features would be created.

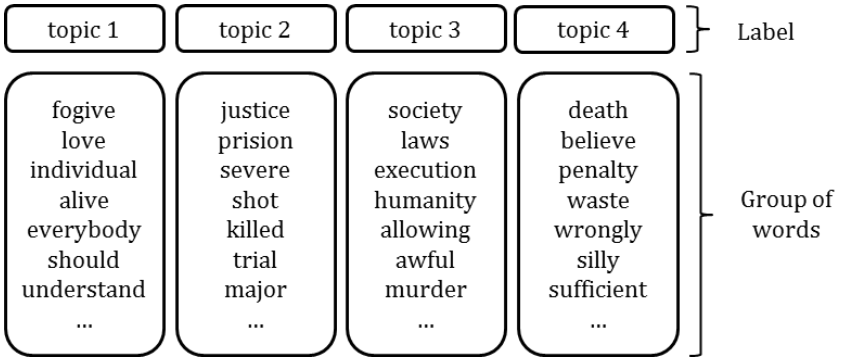

Figure 1. Example of generated topics by using LDA.

3.2 Dataset

To conduct experiments with our approach, we use the corpus proposed in the author identification task of PAN 2015 [7]. The dataset consists of four subsets, each set written in different languages: English, Spanish, Dutch and Greek. Subsets have significant differences. The English subset consists of dialog lines from plays; the Spanish subset consists of opinion articles of online newspapers, magazines, blogs and literary essays; the Dutch subset is formed by essays and reviews; and the Greek subset is formed by opinion articles of categories as politics, health, sports among others. The corpus also has different number of documents per subset detailed in

Table 1. In addition, each language consists of several problems to solve which are specifically defined below (Section 3.3).

Due to its nature, this dataset focused on problems which require capturing more specific information about the writing style of the author. For example, suppose we know a person who worked for a newspaper, writing articles about sports; but one day, this person decides to be independent and spend her life writing horror novels. One possible task can be to find out which articles belong to the sport ex-writer among sport articles of different authors—in this case, the vocabulary of the documents can uncover the author; for instance, by her usage rate of n-grams as features. On the other hand, another possible task is to discover whether a horror novel was written by the novelist, based on the sport articles which she wrote before., This is a drastic change in genre and topic of the documents, i.e., the intersection between vocabularies of the documents would be substantially reduced.

Table 1. Specific values for dataset of author identification task 2015

Language	Training problems			Test problems			Kind
	Items	# docs	Avg. words x doc.	Items	# docs	Avg. words x doc.	
English	100	200	366	500	452	536	Cross-topic
Spanish	100	500	954	100	1000	946	Cross-topic/genre
Dutch	100	276	354	165	380	360	Cross-genre
Greek	100	100	678	100	500	756	Cross-topic

3.3 Method

As an attempt to overcome this problem, we propose to use Latent Dirichlet Allocation (LDA) for extracting semantic information of the corpus. As mentioned before, given a collection of texts, LDA is able to find relations between words by their position in the text. Common stylistics approaches try to find discriminating symbols in the documents so they can distinguish between two documents written by different authors; however, as we stated before (Section 2), while texts become shorter, the amount of symbols is not enough to produce effective discriminate features. This fact becomes worse when authors number is increased. In the case of LDA we expect this issue to be less problematic.

We infer that writers have different ways to link words due to the fact that each writer makes use of favorite phrases. For example, some author usually may use the phrase “the data gathered in the study suggests that” in contrast to other author who uses “the data appears to suggest that”. Thus, the words “the, in, to, that” can be included in different topics since, unlike LSA [10]. LDA can assign the same word to

different topics as an attempt to better handle polysemy. As a result, to use several words at different rates shall result in different topic distributions for each document.

The task of the dataset used for this study is as follows. For each language or subset of the dataset there are specific number of problems; for each problem in turn there are from one to five documents considered as known and one document considered as unknown. These known documents are written by the same author. To solve a specific problem, we must find out whether the unknown document was written by the same author which writes the known documents.

To represent each problem, all documents in the dataset are processed with LDA. Then, we obtain vectors (with real values – probability of each topic) which represent known and unknown documents. Based on a specific problem, we do a subtraction between each known-document's vector and the unknown-document's vector (let us remember that there is only one unknown document by problem, however there are from one to five known documents). We found that converting real values to {0, 1} values slightly improved final results, so we used the arithmetic mean as threshold; 0 represents topic absence and 1 topic presence (above a certain threshold). Therefore, the subtraction between vectors can result in two possible values: 0 when topics are equal and 1 when topics are different (See Figure 2).

Figure 2. Example of subtraction between known-document's vector and unknown-document's vector

4 Results

In the following experiments, we use a Naïve Bayes classifier for classification. For all experiments we chose the number of topics to be 3. Therefore, patterns of three features were generated by each document. We found that varying the topics number, changed the performance classification. There is not a method for determining how many topics we should to choose for incrementing performance. Thus, we had to fix an interval until we achieved the best results. We show in Table 2 results of performance measures (explained below) regarding the number of topics selected. This table shows that the best results are around 3 topics.

Interestingly, with vectors with only a few of topics, we obtained over 64% accuracy. Actually, one might suppose that documents could have been categorized

by subject; however, that assumption is unlikely because, as we showed in Section 3.2, the dataset used is formed of topics and genres mixtures.

Table 2. Selection of topics number based on PAN-2015's author identification task measures

No. Topics	c@1	AUC	FS
2	0.228	0.228	0.052
3	**0.856**	0.807	**0.691**
4	0.702	**0.908**	0.637
5	0.774	0.863	0.668
6	0.660	0.695	0.459
7	0.770	0.806	0.621
8	0.684	0.797	0.545
9	0.730	0.753	0.550
10	0.711	0.834	0.593
20	0.488	0.503	0.245
40	0.496	0.505	0.250
60	0.496	0.468	0.232
80	0.524	0.568	0.298
100	0.468	0.497	0.233

We conducted two experiments for knowing whether two documents written by the same author will be similar on their distribution of topics. Figure 3 shows the sum of all differences by topic in the test dataset for English. As we can see, the amount of differences is high when texts are written by different authors. In Figure 4 is also showed that differences for Spanish language.

We classified the dataset without pre-processing and show in Table 3 the following values: Accuracy, F-measure (F), Precision (P), Recall (R). While accuracy is a measure used in many works on deception detection and it provides us a point of comparison with other results, we also opted for showing precision, recall, and F-measure; this allows for a deeper analysis of outputs. Thus, precision shows the percentage of selected texts that are correct, while recall shows the percentage of correct texts that are selected. Finally, F-measure is the combined measure to assess the P/R trade-off.

We classified the dataset without pre-processing and show in Table 3 the following values: Accuracy, F-measure (F), Precision (P), Recall (R). While accuracy is a measure used in many works on deception detection and it provides us a point of comparison with other results, we also opted for showing precision, recall, and F-measure; this allows for a deeper analysis of outputs. Thus, precision shows the percentage of selected texts that are correct, while recall shows the percentage of

correct texts that are selected. Finally, F-measure is the combined measure to assess the P/R trade-off.

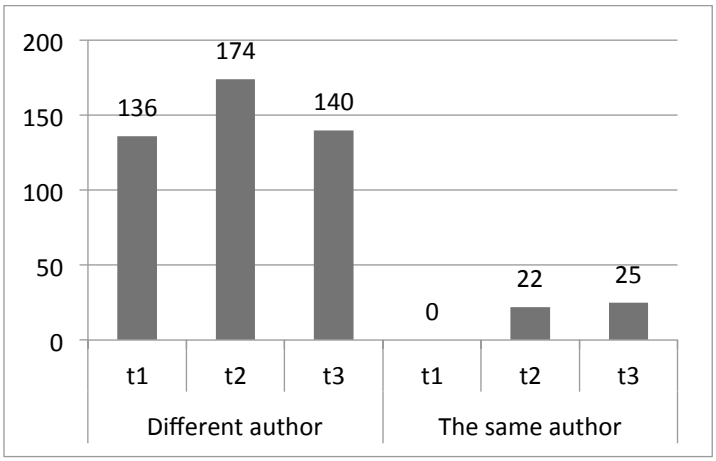

Figure 3. Topic differences between document written either the same or different author (English subset)

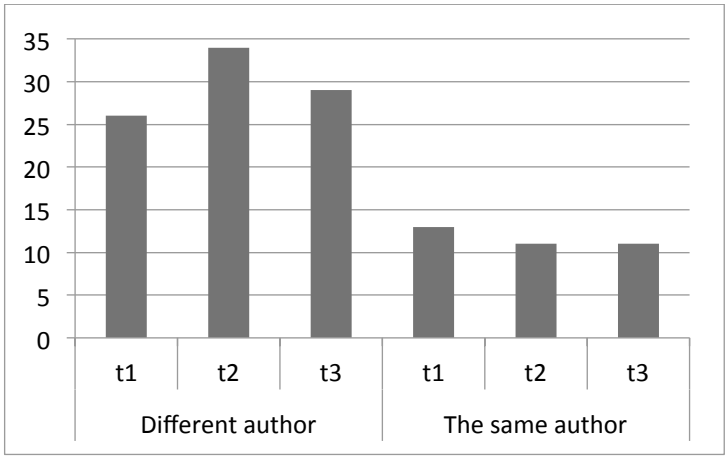

Figure 4. Topic differences between document written either the same or different author (Spanish subset)

We obtained the best result for English subset with 85.6% accuracy even when it has the biggest training set (500 problems) of the corpus. Spanish subset ranks second with 76.0% accuracy, Dutch subset reached 70.9% accuracy and finally Greek subset reached 64.0% accuracy. Both English and Greek subsets obtained the first and the last place of the results (Table 3) respectively; therefore, we cannot infer that the topics mixture made the difference in results since both subsets consist of themes mixture and one of them was not affected. Similarly, for both Spanish and Dutch

subsets (second and third place respectively), results did not lead to conclude that the genre mixture had some correlation on it. For these reasons, we consider that the results were directly affected by the training and test set's document selection and not by the type of text.

Table 3. Results of each subset classification

Subset	Accuracy (%)	Precision	Recall	F-measure
English	85.6	0.864	0.856	0.855
Spanish	76.0	0.760	0.760	0.760
Dutch	70.9	0.733	0.709	0.702
Greek	64.0	0.646	0.640	0.640

We compare our results with those obtained in author identification task at PAN 2015 evaluation lab [7]. Therefore, we calculated, as PAN-2015 task's authors, a final score which is the product of two values: c@1 [11] and area under the ROC curve (AUC) [12]. The former is an extension of the accuracy metric and the latter is a measure of classification performance which provides more robust results than accuracy.

We show in Table 4 the better results obtained for each language subset by participants of PAN-2015 task. According to those results, our method seems to perform well for both English and Dutch languages. This work outperforms FS results with regard to English subset and had better performance than Bartoli et al. and Bagnall's result with regard to Dutch subset. On the other hand, for both Spanish and Greek subsets the proposed method did not show good performance however, ROC curve results showed that predictions are acceptable.

Table 4. Results comparison with other authors. FS=c@1*AUC.

Author	Measure	Subset			
		English	Spanish	Dutch	Greek
Bagnall 2015	c@1	0.757	0.814	0.644	0.851
	AUC	0.811	0.886	0.700	0.882
	FS	0.614	0.721	0.451	**0.750**
Bartoli et al. 2015	c@1	0.559	0.830	0.689	0.657
	AUC	0.578	0.932	0.751	0.698
	FS	0.323	**0.773**	0.518	0.458
Moreau et al. 2015	c@1	0.638	0.755	0.770	0.781
	AUC	0.709	0.853	0.825	0.887
	FS	0.453	0.661	**0.635**	0.693
This work	c@1	0.856	0.760	0.709	0.640
	AUC	0.808	0.737	0.785	0.688
	FS	**0.692**	0.560	**0.556**	0.440

5 Conclusions

A common approach to verify authorship is attempting to find the author's writing style. Therefore, the assumption is that by using that approach, it is possible to capture specific features to discriminate one author from others. This hypothesis is hard to prove; nevertheless, it is known that certain amount of data is necessary to find more appropriate features leading to high classification performance. Data is a problem, for instance, for the forensic field, since hardly there are long texts and they are in different domains. We showed in this work how LDA responds to verify authorship when there is limited data; i.e., only from one to five short texts written by a specific author to determine whether an unknown document belongs to the same author. Furthermore, the used datasets consist of topic and genre mixtures.

Basically, we used documents distributions to capture what we call the authors' fingerprint. Then, by subtraction between topic distributions, we found that documents written by different author tend to differ more than those written by the same author. This approach allowed us to achieve 74% accuracy on average.

References

[1] Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), 538-556.

[2] Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., ... & Barrón-Cedeño, A. (2014, September). Overview of the Author Identification Task at PAN 2014. In *CLEF (Working Notes)* (pp. 877-897).

[3] Nirkhi, S., & Dharaskar, R. V. (2013). Comparative study of authorship identification techniques for cyber forensics analysis. *arXiv preprint arXiv:1401.6118*.

[4] Layton, R., Watters, P., & Dazeley, R. (2013). Local n-grams for Author Identification. Notebook for PAN at CLEF.

[5] Bergsma, S., Post, M., & Yarowsky, D. (2012, June). Stylometric analysis of scientific articles. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 327-337). Association for Computational Linguistics.

[6] Bradley, J. K., Kelley, P. G., & Roth, A. (2008). Author identification from citations. Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep.

[7] Stamatatos, E., Daelemans, W., Verhoeven B., Juola, P., López-López A., Potthast, M., Stein, B. (2015, September). Overview of the Author Identification Task at PAN 2015. In *CLEF (Working Notes)*

[8] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[9] Verhoeven, B., & Daelemans, W. (2014). CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *LREC* (pp. 3081-3085).
[10] Dumais, S. T. (2004). Latent semantic analysis. Annual review of information science and technology, 38(1), 188-230.
[11] Peñas, A., & Rodrigo, A. (2011, June). A simple measure to assess non-response. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 1415-1424). Association for Computational Linguistics.
[12] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
[13] Pimas, O., Kröll, M., & Kern, R. (2015). Know-Center at PAN 2015 author identification. Working Notes Papers of the CLEF.
[14] Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012, May). On the feasibility of internet-scale author identification. In 2012 IEEE Symposium on Security and Privacy (pp. 300-314). IEEE.
[15] Pateriya, P. K. (2012). A Study on Author Identification through Stylometry. *International Journal of Computer Science & Communication Networks*, *2*(6), 653.
[16] Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005, June). Author identification on the large scale. In Proc. of the Meeting of the Classification Society of North America (p. 13).
[17] Castro, A., & Lindauer, B. (2012). Author Identification on Twitter.
[18] Pavelec, D., Justino, E., & Oliveira, L. S. (2007). Author identification using stylometric features. Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, 11(36), 59-66.
[19] Green, R. M., & Sheppard, J. W. (2013, May). Comparing Frequency-and Style-Based Features for Twitter Author Identification. In *FLAIRS Conference*.
[20] Afroz, S., Brennan, M., & Greenstadt, R. (2012, May). Detecting hoaxes, frauds, and deception in writing style online. In 2012 IEEE Symposium on Security and Privacy (pp. 461-475). IEEE.