ISSN 2007-9737

Author Verification Using a Semantic Space Model

Ángel Hernández-Castañeda[1,2], Hiram Calvo[1]

[1] Instituto Politécnico Nacional (IPN),
Centro de Investigación en Computación (CIC),
Mexico

[2] Tecnológico de Estudios Superiores de Tianguistenco (TEST),
Tecnológico Nacional de México (TecNM), Estado de México,
Mexico

ahernandez_a12@sagitario.cic.ipn.mx, hcalvo@cic.ipn.mx

Abstract. In this work we propose to solve the author verification problem using a semantic space model through Latent Dirichlet Allocation (LDA). We experiment with the corpus used in the author identification tasks at PAN 2014 and PAN 2015. These datasets consist of subsets in the following languages: English, Spanish, Dutch and Greek. Each problem contained in these corpora is formed by one to five known documents which were written by one author and one unknown document. The task is to predict whether the unknown document was written by the author who wrote the known documents. We processed the documents in the dataset and captured the fingerprint of authors by generating a probabilistic distribution of words in the documents. In PAN 2015 classification, we achieved 81.6%, 75.4%, 74.1%, 67.1% accuracy for each English, Spanish, Dutch and Greek subset respectively. In particular for the English subset, we outreached the best result reported in both competitions.

Keywords. Author verification, semantic space model, cross-genre, cross-topic, latent Dirichlet allocation.

1 Introduction

Author verification is an important problem to solve, since many tasks require recognizing the author who wrote a specific text. For example, knowing which author wrote an anonymous book, or identifying notes of a serial killer. In this paper we deal with an author verification challenge in a more realistic setting.

Specifically, datasets used (two of them) consists of one to five documents by a known author and one document by an unknown author. Datasets are formed by 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 which wrote the known texts. It is important to note that this task becomes more difficult when the dataset is composed of short documents, since common approaches are not able to capture effective models with few amounts of words [22]. However, in real cases within the forensic field, long texts rarely exist.

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

In this work we use semantic information to find features that help to discriminate texts. For this purpose, we create a model using Latent Dirichlet Allocation. By using this method, we take into account all vocabulary from all texts at the same time, and after a statistical process, find to what extent the relations between words are given in each document.

LDA is a statistical algorithm that considers a text collection as a topics mixture. 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. Then we use machine learning algorithms to classify the obtained patterns.

We evaluate the proposed approach on two datasets where an author verification task is tackled: the corpus PAN 2014 and PAN 2015.

2 Related Work

Several works have attempted to study the authorship identification challenge by generating different kinds of features [10, 12]. The nature of each dataset can determine the difficulty of the task, that is, how hard it will be to extract appropriate features [1, 8]. In [13], a study of techniques can be found, which show that, while the number of authors increases and the size of training dataset decreases, the classification performance decreases.

This seems logical, since the size of training data is smaller, and thus, the identification of helpful features is affected.Many works address author identification through their writing style [15, 16, 25]. For instance, in [9], style-based features are compared to the BoW (Bag of Words) method. They attempt to discriminate authors from texts in the same domain obtained from Twitter.

Style markers such as characters, long words, whitespaces, punctuation, hyperlinks, or parts of speech, among others, are included. The authors found that a style-based approach was more informative than a BoW-based method; however, their best results were obtained by 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 another, but not from every other author.

Stylistic features can be also applied to other tasks. In [2], Bergsma et al. combined features to address two-class problems. They attempted to obtain style, Bag of Words (BoW) and syntax features to classify native and non-native English writers, texts written for conferences or workshops (classification of venue) and texts written by male or female writers (classification of gender).

Their dataset consists of texts that are scientific articles—this kind of texts is more extensive, unlike e-mail, tweets, and other short texts. So, this could have led to identify non-native written texts with promising accuracy. Nevertheless, long texts do not ensure good results, since their classification tasks on venue and gender obtained low accuracy.

The purpose of identifying authorship can vary. For example, Bradley et al. [4] attempt to prove that it is possible to find out which author wrote an unpublished paper (for conference or journal); only by considering the cited works in it. By using latent semantic analysis (LSA) [6], the authors propose to create a term-document matrix wherein possible authors are considered as documents while cited authors are considered as terms. LSA assumes that words that are close in meaning will occur in similar pieces of text. The results of Bradley et al. showed that the blind review system should be questioned.

Another example is Castro and Lindauer [5], with the task of finding out whether Twitter user's identity can be uncovered by their writing style. The authors focused in features as word shape, word length, character frequencies, and stop words' frequencies, among others. With an RLSC (regularized least square classification) algorithm, the authors correctly classified 41% of the tweets.

In Pimas [18] the author verification task is addressed by generating three kinds of features: stylometric, grammatical and statistical. Pimas et al. study is based on the PAN 2015 authorship verification challenge.

In that work, topics distribution is considered as well, but they argue against using it because the dataset is formed of topic mixtures in a way that affected their results. A cross validation model (10 folds) shows good performance; on the other hand, the model showed overfitting using the specified training and test sets.

3 Proposed Approach

As we found in the previous section, features based on stylistics, syntactic and lexical information consider separately each written

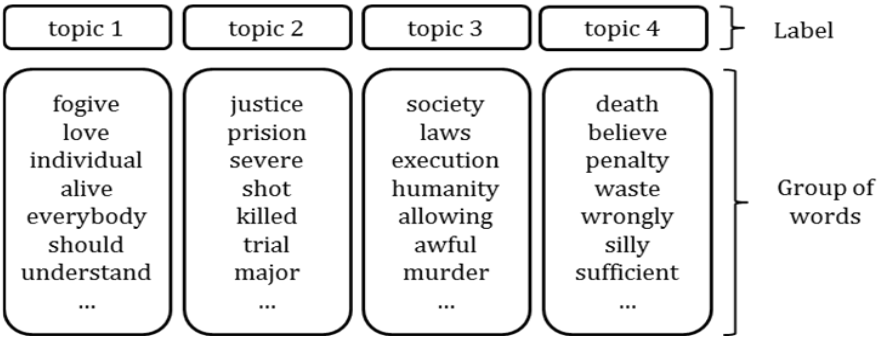

Fig. 1. Example of generated topics by using LDA

document. For instance, it is possible to obtain stylistic features by counting the number of stop words and characters by document.

These occurrences of symbols, in short and cross-domain texts, can merely be random and they do not allow capturing writing style.

In a collection of cross-domain texts, we infer that linked words and the distribution of them in the texts may provide more informative features, since, in first instance, LDA's estimated distributions of topics will depend on the content of each text in the collection.

In this section we present our method for generating features. First, in Section 3.1 we detail the source of features we use. Next, in Section 3.2 we describe the datasets used in this work for evaluation, and finally in Section 3.3 we give details on our feature vector construction.

3.1 Source of Features

Latent Dirichlet Allocation [3] is a probabilistic generative model for discrete data collections such as a collection of texts; 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 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-engineer* this process by finding the most likely set of topics, which may compose a document. LDA automatically generates the groups of words (topics); see Figure 1.

Accordingly, LDA can infer, given a fixed number of topics, how likely that is each topic (set of words) appears in a specific document of a collection. For example, in a collection of documents and 3 latent topics generated with the LDA algorithm, each document would have different distributions of 3 likely topics. That also means that vectors of 3 features would be created.

3.2 Datasets

The proposed approach was tested with different datasets. Included datasets are PAN 2015, which is based on cross-topic and cross-genre documents, and PAN 2014, consisting of specific-domain documents.

Table 1. Dataset details for the PAN 2015 task of author identification

Language	Training problems			Test problems			Type
	Problems	No. docs	Avg. words per doc.	Problems	No. docs	Avg. words per 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

3.2.1 PAN 2015 Corpus

To conduct experiments with our approach, we used the corpus proposed in the author identification task of PAN 2015 [23]. 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 on dialog lines from plays; the Spanish subset consists on 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 amount of documents per subset, as detailed in Table 1. In addition, each language consists of a number of problems to solve which are specifically defined below (Section 3.3).

Due to its nature, this dataset focused on problems that require capturing more specific information about the way of writing of the author. For example, suppose we know a person who worked for newspaper writing articles about sports; nevertheless, that person decides to be independent and spends life writing horror novels.

One possible task could 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 his n-grams-usage rate, this kind of task is called **cross-topic**.

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 or he wrote before. Consequently, there is a drastic change in genre and topic of the documents, i.e., the intersection between vocabularies of the documents would be substantially reduced. This is called **cross-genre**.

3.2.2 PAN 2014 Corpus

The PAN 2014 corpus [24], like PAN 2015, consists of documents written in four languages: Dutch, English, Greek, and Spanish. However, PAN 2014 documents are in the same domain, that is, they do not contain mixtures of genre. Each dataset contains one up to five known documents and one unknown document; the challenge is to find out whether the unknown document was written by the same author who wrote the known documents. We show in Table 2 details about the PAN 2014 corpus.

3.3 Method

There are different works that have used LDA as source of features. For example, Pacheco et al. [14] faced author verification challenge of PAN 2015 by proposing a scheme based on the universal background model [19]. By that scheme three feature vectors were created: a vector for author, a vector for each set of known documents and a vector for each unknown document. The three vectors are encoded to produce a new vector. Then, features (including those generated by LDA) were selected by using a random forest which hierarchically determines the importance of each feature.

In the same way, Moreau et al. [11] used a genetic algorithm for selecting the best sources of features (strategies) for author verification of PAN 2015. One of the strategies was LDA, with which two vectors of five topics were generated by

Table 2. Dataset details for the PAN 2014 task of author identification

Language	Training problems			Test problems			Genre
	Problems	No. docs	Avg. words per doc.	Problems	No. docs	Avg. words per doc.	
Dutch	96	268	412.4	96	287	398.1	Essays
Dutch	100	202	112.3	100	202	116.3	Reviews
English	200	729	848.0	200	718	833.2	Essays
English	100	200	3,137.8	200	400	6,104.0	Novels
Greek	100	385	1,404.0	100	368	1,536.6	Articles
Spanish	100	600	1,135.6	100	600	1,121.4	Articles

executing the following steps: first, the authors split documents on character n-grams; next, a set of five topics for each author's text is sought; finally, their algorithm tries to find a set of five topics for each document to be verified (unknown document). With these two vectors the authors obtain a measure that indicates how similar the vectors are, with regard to the style; that measure is used as fitness for a genetic algorithm.

Savoy [20], in a way similar to us, uses distribution of topics calculated by LDA. The challenge is to find out which author, of a set of authors, wrote a specific document. The symmetric Kullback-Leibler distance between the distribution of the unknown document and the distribution of the documents of the authors was calculated.

Then, the author is assigned with regard to the closest distance. This task is called authorship attribution due to the fact that the challenge is to identify the authors of anonymous text. The authorship attribution task was also studied by Seroussi et al. [21]; in their research topic distributions were viewed as vectors of features (patterns), which were directly classified.

The main difference between other works mentioned above is that, in this work we generate features by using LDA under the assumption that one specific author has a specific way of relating words, and thus generate a text.

That assumption led us to analyze which words are linked in a written text, and to which extent the set of linked words are present in the text. With this information we subtract document vectors to create new vectors that form a training set; this process is detailed below.

Specifically, we propose to use Latent Dirichlet Allocation for extracting semantic information from the corpus. As mentioned before, given a collection of texts, LDA is able to find relations between words based on the way they are used in the text. On the other hand, common stylistics approaches use symbol rates in the documents for distinguishing between two documents written by different authors.

As we stated before (Section 2), while texts become shorter, the amount of symbols tend to be not enough to produce effective discriminating features. This worsens when the number of authors is increased.

We infer that writers have different ways of linking words due to the fact that each writer makes use of specific phrases. In addition, they use words at different rates, and thus, written texts keep a background structure.

For example, some author usually may use the phrase "the data gathered in the study suggests that..." in contrast to another author who uses "the data appears to suggest that." We can see that the words "the, in, to, that" can be included in different topics since, unlike LSA [6], LDA can assign the same word to different topics to better handle polysemy.

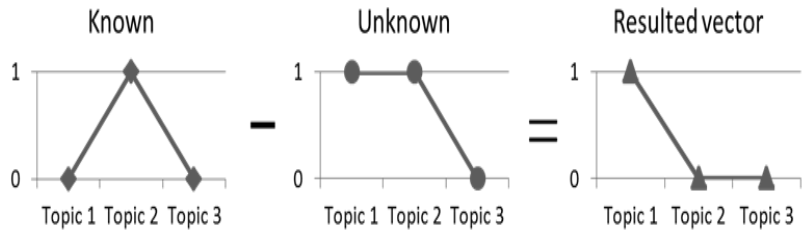

Fig. 2. Example of subtraction between known-document's vector and unknown-document's vector

As a result, to use those 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 is a specific number of problems; for each problem in turn, there are 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 have to find out whether the unknown document was written by the same author who wrote 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 one to five known documents).

We found that converting the real values to {0, 1} values slightly improved final results, so we binarized them using the arithmetic mean as threshold; 0 represents topic absence and 1 represents the presence of the topic. Therefore, the subtraction between vectors resulted in two possible values: 0 when topics are equal and 1 when topics are different (see Fig.).

4 Results

In the following experiments we used Naïve Bayes for classification. In addition, different *n*-topics for LDA were specified. Therefore, patterns of *n* features were generated for each document. We found that varying the topics number also changed the performance of classification. There is no a priori method for determining how many topics we should choose for incrementing performance; thus, we have to fix an interval until we achieve the best results.

Due to the fact that LDA is a stochastic method, the obtained result for each experiment can be different, so we show the average of 100 experiments for each number of topics tested. In addition, standard deviation is shown. Note that each average of 100 experiments was calculated independently for each measure; that is, product of accuracy and ROC area, on the corresponding tables, will not be proportional to FS (final score) measure.

The FS measure is the product of two values: c@1 [17] and the area under the ROC curve (AUC) [7]. The former is an extension of the accuracy metric and the latter is a measure of classification performance that provides more robust results than accuracy.

Under the assumption that in a real case the test set is unknown, a five-cross validation on the training set was done to find the optimal number of topics, and then we used that parameter to evaluate on the test sets.

4.1 Results of PAN 2015 Classification

We show in Table 3, Table 4, Table 5 and Table 6 a detailed analysis of the classification performance of each language on author identification PAN 2015. As can be seen, we can reach the best performance setting different

Table 3. Performance obtained for English, setting different number of topics

Topics	Binary values				Original values			
	c@1	ROC area	FS	FS-SD	c@1	ROC area	FS	FS-SD
2	0.743	0.733	0.554	0.129	0.495	0.407	0.202	0.019
3	**0.816**	0.853	0.697	0.041	0.672	0.699	0.474	0.090
4	0.800	**0.847**	**0.679**	0.049	0.633	0.637	0.407	0.080
5	0.781	0.835	0.654	0.067	0.631	0.646	0.412	0.085
6	0.759	0.811	0.618	0.075	0.622	0.621	0.389	0.064
7	0.778	0.790	0.586	0.084	0.595	0.598	0.358	0.064
8	0.746	0.800	0.599	0.064	0.591	0.594	0.356	0.083
9	0.735	0.787	0.581	0.075	0.616	0.624	0.388	0.073
10	0.724	0.767	0.559	0.084	0.604	0.603	0.367	0.067

Table 4. Performance obtained for Spanish, setting different number of topics

Topics	Binary values				Original values			
	c@1	ROC area	FS	FS-SD	c@1	ROC area	FS	FS-SD
2	0.633	0.632	0.402	0.037	0.637	0.672	0.428	0.032
3	0.730	0.765	0.561	0.080	0.673	0.702	0.473	0.046
4	**0.750**	**0.783**	**0.589**	0.066	0.678	0.697	0.474	0.053
5	0.740	0.776	0.576	0.071	0.678	0.698	0.475	0.047
6	0.751	0.777	0.586	0.072	0.661	0.694	0.463	0.061
7	0.754	0.777	0.589	0.075	0.664	0.697	0.465	0.068
8	0.726	0.756	0.551	0.072	0.646	0.676	0.439	0.064
9	0.719	0.747	0.540	0.079	0.639	0.667	0.429	0.070
10	0.715	0.738	0.530	0.075	0.635	0.658	0.420	0.066

topics. For instance, the method achieves the best result in the case of English when we set four topics (see Table 3).

For each table, accuracy, ROC area, FS, and FS-SD (final score – standard deviation) are shown. Furthermore, a comparison between the performance obtained with original values and binary values of LDA is shown; this comparison suggests that, at least for the test environment in this study, binary values obtain better performance on classification. It is interesting to note that with vectors formed for few topics (one up to ten), we are able to obtain over 64% accuracy.

As a matter of fact, one may suppose that documents could have been categorized by subject; however, that assumption seems unlikely, due to the fact that, as we showed in Section 0, the used dataset is conformed by a mixture of subjects. We conducted two experiments aiming to find whether two documents written by the same author will be similar based on their distribution of topics.

Fig. 3 shows the sum of all differences by topic in the test dataset for English. As can be seen, the number of differences is high when texts are written by different authors.

In Fig. 4 the same differences are shown for the Spanish language. We classified the dataset without pre-processing and found the following values shown in table 7: Accuracy, F-measure (F), Precision (P), Recall (R). In this table we

Table 5. Performance obtained for Dutch, setting different number of topics

Topics	Binary values				Original values			
	c@1	ROC area	FS	FS-SD	c@1	ROC area	FS	FS-SD
2	0.722	0.721	0.521	0.030	0.682	0.703	0.479	0.023
3	**0.741**	**0.751**	**0.557**	0.030	0.685	0.710	0.487	0.029
4	0.717	0.736	0.530	0.067	0.671	0.707	0.475	0.035
5	0.675	0.730	0.496	0.067	0.636	0.676	0.432	0.050
6	0.668	0.724	0.485	0.049	0.617	0.661	0.409	0.044
7	0.681	0.730	0.498	0.043	0.619	0.651	0.404	0.046
8	0.690	0.730	0.505	0.041	0.626	0.653	0.410	0.045
9	0.696	0.727	0.508	0.050	0.633	0.660	0.419	0.040
10	0.690	0.715	0.494	0.045	0.629	0.657	0.414	0.038

Table 6. Performance obtained for Greek, setting different number of topics

Topics	Binary values				Original values			
	c@1	ROC area	FS	FS-SD	c@1	ROC area	FS	FS-SD
2	0.616	0.616	0.384	0.079	0.594	0.611	0.364	0.054
3	0.643	0.674	0.440	0.090	0.564	0.591	0.336	0.060
4	0.668	0.695	0.469	0.087	0.564	0.592	0.336	0.053
5	0.665	0.696	0.466	0.077	0.561	0.599	0.338	0.059
6	**0.671**	**0.709**	**0.479**	0.086	0.567	0.600	0.344	0.075
7	0.661	0.696	0.468	0.114	0.566	0.601	0.344	0.074
8	0.664	0.703	0.471	0.089	0.561	0.595	0.337	0.066
9	0.654	0.690	0.455	0.084	0.585	0.603	0.356	0.071
10	0.669	0.703	0.476	0.100	0.593	0.612	0.366	0.071

Table 7. Results of each subset classification. Accuracy (Acc), Precision (P), Recall (R) and F-measure (F)

Subset	Type	Acc	P	R	F
English	Cross-topic	85.6	0.864	0.856	0.855
Spanish	Cross-topic/genre	76.0	0.760	0.760	0.760
Dutch	Cross-genre	70.9	0.733	0.709	0.702
Greek	Cross-topic	64.0	0.646	0.640	0.640

show values corresponding to the experiment closest to the average result for each language.

While accuracy is a measure used in many works on author identification and provides a point of comparison with other results, we also opted for showing precision, recall, and F-measure; this allows for a deeper analysis of results: precision shows the percentage of selected texts that are correct, while recall shows the percentage of correct texts that are selected.

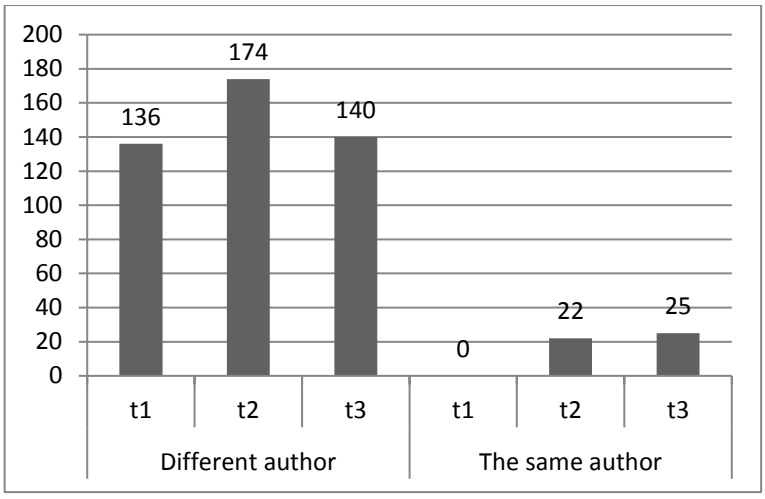

Fig. 3. Topic differences between documents written either by the same or by different author (English subset)

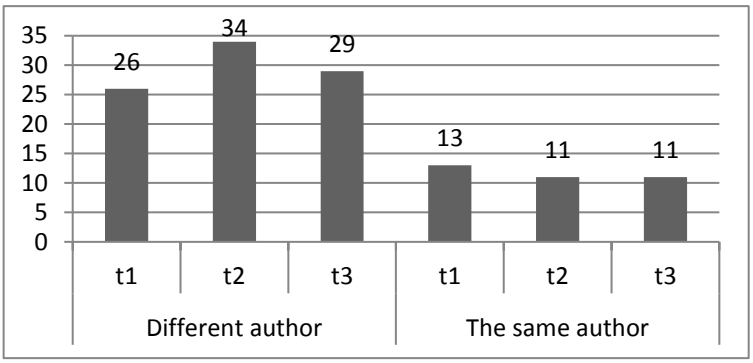

Fig.4. Topic differences between documents written either by the same or by different author (Spanish subset)

Finally, F-measure is the combined measure to assess the P/R trade-off.

According to Table 7, we obtained the best result for the English subset with 85.6% accuracy, even when it has the largest testing set (500 problems) of the corpus. The Spanish subset ranks second with 76.0% accuracy, the Dutch subset reached 70.9% accuracy and finally the Greek subset reached 64.0% accuracy.

We obtained our best and worst performance with the English and Greek subsets respectively; therefore, we cannot infer that the cross-topic setup made the difference in results since actually both subsets are of the same type (see Table 7) and one of them was not affected. Similarly, for both Spanish and Dutch subsets (second and third place respectively), results do not lead to conclude that the genre mixture has some correlation with them. We compared our results with those obtained in author identification task at PAN 2015 [23]. Therefore, we calculated the same as PAN-2015 task's authors, a final score.

We show in Table 8 the best results obtained for each language subset by participants of the PAN-2015 task. In addition, the worst FS scores of PAN-2015's participants are shown. According to those results, our method seems to perform better for both English and Dutch languages.

Additionally, our method is able to outperform FS results with regard to the English subset and has better performance than Bartoli et al. and Bagnall's result with regard to the Dutch subset.

Table 8. 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	0.816	0.750	0.741	0.671
	AUC	0.853	0.783	0.751	0.709
	FS	**0.697**	0.589	0.557	0.479
Worst result PAN'15	FS	0.201	0.095	0.089	0.212

On the other hand, for both Spanish and Greek subsets the proposed method did not show good performance; however, our results are not among the worst FS scores of PAN-2015's participants and ROC curve results show that predictions are acceptable.

We exclude that the bad results may caused by the length of vocabulary since we counted around 1000 and 600 types on Spanish and Greek documents respectively; and just around 300 and 200 types on English and Dutch documents.

Thus, if the performace depended on the size of vocabulary, better results would be obtained on Spanish and Greek documents. We infer that results depend on the documents source.

That is, English articles were taken from plays, containing dialogs; therefore, those documents may keep certain similar latent structure since all of them were written for a specific purpose. Unlike English documents, the other documents have different purposes; for example, Spanish documents are taken from newspapers and even from personal blogs.

4.2 Results of PAN 2014 Classification

The task of PAN 2014 is very similar to the PAN 2015 task explained above; however there is a main difference with regard to the dataset: in PAN 2014 there is no merging of domains. Thus, it contains six subsets, and each subset consists of texts of the same genre (for example, essays, articles, novels).

In this section, we show the results of PAN 2014 classification. Table 9 and Table 10 show detailed results for the subsets where the PAN 2014 baseline was reached, and where it was not surpassed, respectively.

In this case, we outperformed the baseline, set by PAN 2014 challenge, for Dutch reviews, English novels and English essays (three of six subsets, see Table 11).

Table 9. Detailed results of classification on genres where baseline was reached

	Dutch reviews		English novels		English essays	
Topics	**FS**	**FS-DS**	**FS**	**FS-DS**	**FS**	**FS-DS**
10	0.315	0.041	**0.391**	0.062	0.517	0.042
20	0.312	0.058	0.358	0.051	0.533	0.044
30	**0.325**	0.054	0.333	0.030	0.539	0.040
40	0.301	0.048	0.327	0.054	0.535	0.040
50	0.310	0.066	0.296	0.045	0.540	0.043
60	0.283	0.052	0.292	0.041	**0.551**	0.039
70	0.261	0.050	0.279	0.041	0.540	0.032
80	0.256	0.052	0.282	0.043	0.535	0.024
90	0.281	0.046	0.270	0.038	0.536	0.028
100	0.296	0.051	0.268	0.048	0.546	0.034

Table 10. Detailed results of classification on genres where baseline was not reached

	Dutch essays		Greek articles		Spanish articles	
Topics	**FS**	**FS-DS**	**FS**	**FS-DS**	**FS**	**FS-DS**
10	0.661	0.052	**0.436**	0.050	**0.322**	0.047
20	**0.667**	0.045	0.406	0.064	0.310	0.056
30	0.648	0.053	0.360	0.046	0.280	0.050
40	0.628	0.053	0.352	0.060	0.277	0.054
50	0.595	0.037	0.339	0.042	0.280	0.044
60	0.565	0.059	0.329	0.058	0.261	0.046
70	0.548	0.055	0.304	0.052	0.255	0.045
80	0.544	0.064	0.298	0.039	0.259	0.044
90	0.494	0.057	0.288	0.051	0.275	0.061
100	0.475	0.053	0.267	0.045	0.261	0.047

Table 12 shows the best scores achieved by the PAN-2014's participants. Our method only outperforms on English essays classification with regard to the best scores.

In order to find why other languages had a lower performance, we manually analyzed randomly selected Spanish texts where the proposed method yielded bad classification results; we manually looked for cues to find out a consistent characteristic that lead to a missclasification. Nonetheless, texts did not show spelling errors, Spanish slang, or other signs which could help us to differ correctly from incorrectly classified texts. We consider a deeper analysis in a future work is necessary.

5 Conclusions

A common approach to verify authorship is by attempting to model the author's writing style. The assumption is that, by using that approach, it is possible to capture specific features to discriminate one author from others.

That hypothesis is hard to prove; nevertheless it is known that certain amount of data is

Table 11. Performance comparison of this work and PAN-2014 baselines

Genre	c@1	ROC area	FS of this work	FS baseline of PAN 2014
Dutch essays	0.778	0.855	0.667	**0.685**
Dutch reviews	0.560	0.577	**0.325**	0.322
English essays	0.705	0.780	**0.551**	0.288
English novels	0.605	0.644	**0.391**	0.202
Greek articles	0.631	0.686	0.436	**0.452**
Spanish articles	0.555	0.576	0.322	**0.378**

Table12. Score comparison of this work and PAN-2014's participants

Genre	Best score PAN'14	Worst score PAN'14	This work
Dutch essays	0.823	0.307	0.667
Dutch reviews	0.525	0.170	0.325
English essays	0.513	0.270	**0.551**
English novels	0.508	0.225	0.391
Greek articles	0.720	0.281	0.436
Spanish articles	0.698	0.248	0.322

necessary to find more appropriate features leading to a high classification performance.

Finding suitable data is a problem, for instance, when we are talking about forensic field, since there are hardly long texts available and they are in different domains.

We showed in this work how LDA aids 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.

Basically we used document distributions to capture what we call the author's fingerprint. Then, by subtraction between topic distributions, we found that documents written by different authors tend have different fingerprints compared to those written by the same author.

Due to the fact that LDA is a stochastic method, it is necessary to preserve consistency on subtractions; thus, we have to process all documents at the same time. For instance, if we process the training set and after the test set, the topic *x* from the distribution of the document *z* may not correspond with the topic *x* of the distribution of the document *w*. Therefore, in a real case, to classify a new unknown document, it would be necessary to re-process all documents including the new ones.

This approach allowed us to achieve 74% accuracy on average for all different languages included in PAN 2015, and 63.9% accuracy in average on PAN 2014. In both editions, we were able to surpass the best results reported for the English author identification task. Finding a specific reason of performance decrease for other languages has been left as a future work.

Acknowledgements

We thank Instituto Politécnico Nacional (SIP, COFAA and BEIFI), CONACYT, TEST-TecNM, and Red TTL for their support.

References

1. **Afroz, S., Brennan, M., & Greenstadt, R. (2012).** Detecting hoaxes, frauds, and deception in writing style online. *IEEE Symposium on Security and Privacy*, pp. 461–475. DOI: 10.1109/SP.2012.34.

2. **Bergsma, S., Post, M., & Yarowsky, D. (2012).** Stylometric analysis of scientific articles. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp. 327–337, Association for Computational Linguistics.
3. **Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003).** Latent Dirichlet allocation. *Journal of machine Learning research*, Vol. 3, pp. 993–1022.
4. **Bradley, J.K., Kelley, P.G., & Roth, A. (2008).** *Author identification from citations*. Dept. Computing Science, Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep.
5. **Castro, A. & Lindauer, B. (2012).** *Author Identification on Twitter.* Semanticscholar.org.
6. **Dumais, S.T. (2004).** Latent semantic analysis. *Annual review of information science and technology*, Vol. 38, No. 1, pp.188–230.
7. **Fawcett, T. (2006).** An introduction to ROC analysis. *Pattern recognition letters*, Vol. 27, No. 8, pp. 861–874.
8. **Green, R.M. & Sheppard, J.W. (2013).** Comparing Frequency-and Style-Based Features for Twitter Author Identification. *The Twenty-Sixth International FLAIRS Conference*.
9. **Layton, R., Watters, P., & Dazeley, R. (2013).** Local n-grams for Author Identification. *Notebook for PAN at CLEF.*
10. **Madigan, D., Genkin, A., Lewis, D.D., Argamon, S., Fradkin, D., & Ye, L. (2005).** Author identification on the large scale. *Proceedings of the Meeting of the Classification Society of North America,* pp. 13.
11. **Moreau, E., Jayapal, A., Lynch, G., & Vogel, C. (2015).** Author verification: basic stacked generalization applied to predictions from a set of heterogeneous learners. *Working Notes Papers of the CLEF*.
12. **Narayanan, A., Paskov, H., Gong, N.Z., Bethencourt, J., Stefanov, E., Shin, E.C.R., & Song, D. (2012).** On the feasibility of internet-scale author identification. *IEEE Symposium on Security and Privacy*, pp. 300–314.
13. **Nirkhi, S. & Dharaskar, R.V. (2013).** *Comparative study of authorship identification techniques for cyber forensics analysis.* ArXiv preprint 1401.6118.
14. **Pacheco, M.L., Fernández, K., & Porco, A. (2015).** Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification. *CLEF Working Notes.*
15. **Pateriya, P.K. (2012).** A Study on Author Identification through Stylometry. *International Journal of Computer Science & Communication Networks*, Vol. 2, No. 6, pp. 653–657.
16. **Pavelec, D., Justino, E., & Oliveira, L.S. (2007).** Author identification using stylometric features. *Revista Iberoamericana de Inteligencia Artificial*, Vol. 11, No. 36, pp. 59–66.
17. **Peñas, A. & Rodrigo, A. (2011).** A simple measure to assess non-response. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* Vol. 1, pp. 1415–1424, Association for Computational Linguistics.
18. **Pimas, O., Kröll, M., & Kern, R. (2015).** Know-Center at PAN author identification. *Working Notes Papers of the CLEF.*
19. **Reynolds, D.A., Quatieri, T.F., & Dunn, R.B. (2000).** Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, Vol. 10, No. 3, pp. 19–41.
20. **Savoy, J. (2013).** Authorship attribution based on a probabilistic topic model. *Information Processing & Management*, Vol. 49, No. 1, pp. 341–354.
21. **Seroussi, Y., Zukerman, I., & Bohnert, F. (2014).** Authorship attribution with topic models. *Computational Linguistics*, Vol. 40, No. 2, pp. 269–310.
22. **Stamatatos, E. (2009).** A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, Vol. 60, No. 3, pp. 538–556.
23. **Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B. (2015).** Overview of the Author Identification Task at PAN. *CLEF Working Notes.*
24. **Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B. (2014).** Overview of the Author Identification Task at PAN. *CLEF Working Notes*, pp. 877–897.
25. **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. *LREC*, pp. 3081–3085.

Article received on 14/11/2016; accepted on 17/03/2017.
Corresponding author is Ángel Hernández-Castañeda.