Intrinsic Author Verification Using Topic Modeling

Nektaria Potha
nekpotha@aegean.gr
University of the Aegean
Karlovassi, Greece

Efstathios Stamatatos
stamatatos@aegean.gr
University of the Aegean
Karlovassi, Greece

ABSTRACT

Author verification is a fundamental task in authorship analysis and associated with important applications in humanities and forensics. In this paper, we propose the use of an *intrinsic profile-based* verification method that is based on latent semantic indexing (LSI). Our proposed approach is easy-to-follow and language independent. Based on experiments using benchmark corpora from the PAN shared task in author verification, we demonstrate that LSI is both more effective and more stable than latent Dirichlet allocation in this task. Moreover, LSI models are able to outperform existing approaches especially when multiple texts of known authorship are available per verification instance and all documents belong to the same thematic area and genre. We also study several feature types and similarity measures to be combined with the proposed topic models.

ACM Reference Format:
Nektaria Potha and Efstathios Stamatatos. 2018. Intrinsic Author Verification Using Topic Modeling. In *SETN '18: 10th Hellenic Conference on Artificial Intelligence, July 9–15, 2018, Rio Patras, Greece.* ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3200947.3201013

1 INTRODUCTION

Due to the rapid growth of online information text categorization has become one of the key techniques providing effective solutions for handling and organizing the huge volumes of electronic text data. The three main directions of distinguishing between texts are their topic, sentiment, and style. The latter is a useful factor to identify document genre and reveal information about the author(s). In recent years, authorship analysis attracts constantly increasing attention [11, 28] with applications closely related with areas such as humanities (e.g., revealing the author of novels published anonymously, verifying the authorship of literary works, etc.) [18, 32] and forensics (e.g., discovering authorship links between proclamations of different terrorist groups, resolving copyright disputes, revealing multiple aliases of the same user in social media, verifying the authorship of suicide notes, etc) [1, 33].

Authorship attribution is considered as the major discipline of authorship analysis in terms of published studies, as it contributes decisively in the revealing of the identity of the author of an unsigned text by measuring some of the text's features and quantifying stylistic choices which represent the writing style of a text [28]. In each authorship attribution case, a set of candidate authors and undisputed sample documents of them are considered. Then, taking into consideration and estimating the stylistic similarities between their personal writing style and that of the anonymous document under investigation, we are looking for the most likely candidate author. The two main forms of this task examined in the relevant literature are the *closed* and *open-set attribution*. In the first, the true author of the questioned documents is certainly included in the set of candidate authors. This is the easiest version of the problem and most studies have focused on this, providing very encouraging results [25, 27, 28]. In the second one, the true author of the questioned documents is not necessarily contained in the set of candidate authors. This is a much more difficult scenario especially when the size of the candidate set is small [15]. This setting fits all kind of applications including cases where everyone can be the true author of a questioned document (e.g., identifying the person behind a post in a blog). *Authorship verification* can be considered as a special case of *open-set attribution* when the candidate set is singleton. It is a fundamental task since any authorship attribution instance can be decomposed into a set of verification instances [16]. So, the ability of a method to deal effectively with this basic task is crucial.

More formally, let D_{known} be a set of documents all written by the same author and d_u a document of disputed authorship. Then, an authorship verification method estimates the probability $p(d_u|D_{known})$ that indicates how likely it is for d_u and D_{known} to be written by the same author. It is essentially a one-class classification task. Author verification methods vary with respect to how they treat the members of D_{known}. According to the *profile-based paradigm* all members of D_{known} are concatenated and a single representation is extracted [28]. This is an author-centric approach. On the other hand, the *instance-based paradigm* treats each member of D_{known} separately. It is a document-centric approach. The majority of existing approaches follow the latter paradigm [29, 30]. From another perspective, *intrinsic* verification methods attempt to estimate the probability score exclusively based on D_{known} while *extrinsic* models also consider external documents (by other authors). The latter approach essentially attempts to transform author verification to a binary classification task and the most successful approaches so far follow this practice [2, 13, 19].

So far, topic modeling features have been used in author verification methods as a complement to other, more powerful features. Typically, Latent Dirichlet Allocation (LDA) features are combined with part-of-speech features [20, 21]. In this paper, we present an *intrinsic profile-based* author verification method that is exclusively based on topic modeling features, more specifically derived by Latent Semantic Indexing. We apply this method to benchmark corpora developed in the framework of PAN-2014 and PAN-2015

SETN '18, July 9–15, 2018, Rio Patras, Greece
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6433-1/18/07…$15.00
https://doi.org/10.1145/3200947.3201013

shared tasks adopting exactly the same evaluation setup. In contrast to current state-of-the-art in this field, we provide evidence that *intrinsic* and *profile-based* authorship verification can be very effective outperforming the top-ranked PAN submissions in most of the cases. Furthermore, we demonstrate that LSI is clearly more effective than LDA in the proposed setting and several parameters and variations of our approach are examined. The rest of this paper is organized as follows: Section 2 presents previous work in authorship verification while Section 3 describes our proposed method and its variations. In Section 4, the performed experiments are presented and Section 5 includes the main conclusions drawn from this study and discusses future work directions.

2 PREVIOUS WORK

The authorship verification task was first discussed by Stamatatos et al. [31]. They used stylometric features provided by an NLP tool and multiple regression to produce the response function for a given verification case. Perhaps, the most well-known early verification method is called *Unmasking* [17]. This method builds an SVM classifier to distinguish between two texts and examines the drop in classification accuracy when the most significant features are gradually removed. This method is especially effective in long documents (e.g., novels) but it is unable to handle relatively short documents (e.g., articles) [24].

From 2013 to 2015, a series of shared tasks relevant to author verification were organized in the framework of PAN evaluation lab [12, 29, 30]. Dozens of submissions were received and evaluated on corpora covering several natural languages and genres. The most successful submissions in these shared tasks used *extrinsic* and *instance-based* approaches [2, 13, 26].

Verification models differ with respect to their view of the task. *Intrinsic* verification models view it as a one-class classification task and are based exclusively on analysing the similarity between D_{known} and d_u. Typical approaches of this category are described in [8, 10, 22]. Such methods are usually robust and fast while they do not require any external resources.

On the other hand, *extrinsic* verification models attempt to transform the verification task to a pair classification task by considering external documents to be used as samples of the negative class. They are usually found to be more effective than *intrinsic* methods [29, 30]. The most influential method of this category called *Impostors* was introduced by Koppel and Winter [19]. This method attempts to decide if d_u is more similar to D_{known} than to D_{ext} (a set of external documents by other authors) building a random subspace ensemble by randomly choosing a subset of features and a subset of external documents in each iteration. The effectiveness of this method depends on the quality of D_{ext} that has to be sampled from a huge and heterogeneous class (i.e., all documents by all other authors). The winners of PAN-2013 and PAN-2014 shared tasks in author verification were modifications of the *Impostors* method [13, 26]. Another recent modification that enhances the information extracted in each repetition of the ensemble is described by Potha and Stamatatos [23].

From another perspective, a set of verification approaches consider each verification problem as an instance of a binary classification task and attempt to train a classifier that can distinguish between positive (same author) and negative (different author) instances [7, 21]. Such methods heavily depend on the properties and representativeness of the training corpus. If such a method is applied to other corpora that have no significant similarities with the training corpus (including the distribution of positive and negative instances), practically they fail [30].

A major outcome of PAN shared tasks was that when multiple verification models are combined in an heterogeneous ensemble, usually the effectiveness is improved in comparison to any individual method [29, 30]. Moreau et al. [20] follow this idea and present very encouraging results even for the most challenging cross-genre cases where the texts within a verification instance differ in genre.

As concerns the stylometric features, most author verification methods use low-level features like word *n*-grams and character *n*-grams avoiding to include more sophisticated (and more noisy) features related to syntactic analysis of texts [29, 30]. A very effective approach that won the PAN-2015 shared task was based on a character-level multi-head recurrent neural network model [2]. So far, topic modeling (LDA) features have been used as a complement to other more powerful features [20, 21]. More recently, the study of Hernandez et al., [9] builds a verification model based completely on LDA features, in an attempt to explore the correlation between words in documents.

3 THE PROPOSED METHOD

The main idea of our proposed approach is to use powerful low-level featuress, like word unigrams and character *n*-grams and apply topic modeling techniques in order to reduce dimensionality and extract more compact and less sparse representations of texts. In a more detailed description, we follow the *profile-based paradigm*, that is all available texts by the same author are concatenated and a single text representation vector is extracted for that author in the reduced latent semantic space (e.g., using LSI). This author vector is then compared with the corresponding vector of a text of disputed authorship and a similarity score is calculated. The highest the value of the similarity score the more likely the disputed text to be written by that author.

More formally, let C be a collection of documents and $\vec{X_d}$ the representation vector of a document d in a high dimensional space (e.g., word or character n-grams). A topic modeling method (e.g., LSI) estimates a latent semantic space based on C and derives a $\vec{X_d}$ representation of each d in that reduced space. Given an author verification instance (d_u, D_{known}), let d_k be the concatenation of all documents in D_{known}. Then, our proposed approach estimates the following score:

$$p\left(d_u|D_{known}\right) \sim similarity\left(\hat{\vec{X}}_{d_u}, \hat{\vec{X}}_{d_k}\right) \tag{1}$$

The training phase of our approach aims at extracting the topic model and the most appropriate parameter values for a specific corpus (i.e., collection of author verification instances with similar properties). It is also important to be noticed that for each corpus separately, topic model is constructed following two options. In the first one, utilizing entirely the training texts. Alternatively, the latter concerns about enriching the documents of the training part of each corpus with additional external documents. A detailed analysis about the construction of topic model and selection of additional documents is reported in section 4.2. In total, the proposed approach has three parameters to be tuned. The first is the feature type (either word unigrams, character 3-grams, 4-grams or 5-grams). The second is the number of latent topics (k). The last one is the similarity function discussed in Section 3.2.

3.1 Topic Models

Topic modeling is one of the most useful processes that contributes decisively to analyze large volumes of data providing a simple way to transform them from a high to low-dimensional representation. Topic modeling techniques achieve to reduce the size of feature space by grouping words of a corpus into relatively few topics and then representing each document as a mixture of these topics. Two of the most well-known and efficiently computable dimensionality reduction techniques are described below:

(1) Latent semantic indexing (LSI) is a popular linear algebraic method that can be used to reduce dimensionality. It attempts to uncover the latent concepts (or structures) in a collection of documents. In other words, LSI is a technique [14] which analyzes relationships between a collection of documents and the terms they contain by producing a set of concepts related to the documents and terms. In the approximation of a small dimensional space LSI can be accomplished by a matrix algebra technique termed Singular Value Decomposition (SVD). If the input Term Frequency matrix is $A= m\times n$, where each row is a document and each column is a term, then the decomposition $A = USV^T$ is called singular decomposition of A, where $U =m\times r$, implying m documents and r concepts and $V= n\times r$, denoting n terms and r concepts, are defined as singular vectors (left and right) of the matrix A. $S=r \times r$ is a diagonal matrix whose entries from upper left to lower right are positive and in decreasing order. These values are the singular values / eigenvalues of A and indicate the strength of the (latent) concept. Aiming to create a new matrix of significantly lower dimension, less sparse in relation to the original matrix A, we reduce the square matrix S retaining the top singular values [4].

(2) Another very popular topic modeling method is Latent Dirichlet Allocation (LDA) [3]. LDA is a generative probabilistic approach where each document is represented as a finite mixture over a set of (latent) topics and each topic is described by a distribution over words. It is closely related to probabilistic latent semantic analysis and the main difference is that LDA assumes that the topic distribution follow a sparse Dirichlet prior. LDA has already been used in previous authorship verification studies [9, 20, 21] and it is used as a baseline in the current study.

Mapping a high dimensional space into a low dimensional space has the effect of reducing noise in the data as well as reducing the sparseness of the original matrix. It is also worth noticing that either LSI or LDA are inherently independent of any language, in consequence of utilizing a strictly mathematical approach. This enables the aforementioned techniques to elicit the semantic content of information written in any language without requiring the use of language-dependent resources.

3.2 Similarity Functions

Given two vectors $\vec{x},\vec{y}$ representing documents x and y and a predefined weight vector $\vec{w}$ which determines the importance of each feature, our task is to estimate how similar x and y are. Cosine similarity is one of the most popular similarity measures applied in information retrieval tasks and is appropriate for high dimensional spaces and sparse representation vectors. Similarity is quantified as the cosine of the angle between vectors. Since each dimension of the vector has a non-negative value, the cosine similarity is non-negative as it is notably used in the positive space. The cosine similarity function of two documents is described as follows:

$$cosine\left(\vec{x},\vec{y}\right) = \frac{\sum_{i=1}^{n} w_i(x_i \cdot y_i)}{\left(\sum_{i=1}^{n}(w_i \cdot x_i^2)\right)^{\frac{1}{2}} \cdot \left(\sum_{i=1}^{n}(w_i \cdot y_i^2)\right)^{\frac{1}{2}}} \tag{2}$$

Furthermore, an alternative and stable similarity metric to compare documents to each other is based on the Euclidean distance. It is the ordinary type of distance in n-dimensional space. Euclidean distance is widely used in text clustering problems and can also support a predefined weights for each dimension. The formula of the Euclidean similarity is defined as:

$$Euclidean\left(\vec{x},\vec{y}\right) = 1 - \left(\sum_{i=1}^{n} w_i\left(x_i - y_i\right)^2\right)^{\frac{1}{2}} \tag{3}$$

Another similarity measure popular in authorship analysis tasks [19] is the Minmax measure:

$$Minmax\left(\vec{x},\vec{y}\right) = \frac{\sum_{i=1}^{n} w_i min\left(x_i, y_i\right)}{\sum_{i=1}^{n} w_i max\left(x_i, y_i\right)} \tag{4}$$

In this study, we consider two forms of the above measures: a simple (or unweighed) version where all features are equally important (i.e., w_i=1 for i=1,...,n) and a weighted version where $\vec{w}$ is provided by the diagonal singular values derived by applying LSI topic modeling.

4 EXPERIMENTS

4.1 Description of Data

We use the authorship verification corpora benchmarks developed in the framework of the relevant PAN shared tasks in 2014 and 2015 covering four languages (Dutch, English, Greek, and Spanish) and genres (newspaper articles, essays, reviews, literary texts, etc.) [29, 30]. Each PAN corpus includes a set of verification problems (instances) and each problem contains a small number (up to 10) texts by the same author and exactly one text of unknown authorship. Each corpus is segmented into a training and an evaluation part and in each case the distribution of positive (same-author) and negative (different-author) instances is balanced.

In PAN-2014 corpora, all texts within a verification problem are in the same language, genre, and thematic area. This ensures that the main differences between texts are due to the personal style of authors. On the other hand, PAN-2015 corpora focus on challenging cross-domain cases where texts within a verification problem are in the same language but the thematic area (PAN-14-EN, PAN-14-GR, PAN-14-SP) or genre (PAN-14-DU) of texts may vary. Each PAN corpus has its own specific properties. For example, in the English corpus of PAN-15 there is exactly one known document per verification problem while in the Spanish corpus of PAN-15 there are exactly four known documents per problem. Moreover, the genre of documents correlate with text-length (e.g., newspaper articles are longer than essays and reviews). All these make comparison of results between different corpora difficult.

The participants of PAN shared tasks were asked to produce a numerical score in [0,1] for each verification problem indicating the probability estimate of a positive answer (the known and unknown texts are by the same author). The evaluation of participants was based on the area under the ROC curve (AUC) [6]. In this paper, we follow exactly the same evaluation settings to achieve comparability of comparison with previously reported results.

4.2 Experimental Setup

To extract a topic model (either LSI or LDA) a set of documents is needed. In this paper, we explore two options. The first one is to use exclusively the available documents from the training part of each corpus [29, 30]. Since in many cases different verification instances share documents, we take under consideration only the unique documents in each training corpus removing duplicate documents. The second option is to enrich the documents of the training part with additional external documents. First, significant terms are extracted from the training documents and several queries are formed and submitted to the Bing search engine. Then, the first results are downloaded. Roughly, 500 external documents are extracted for each PAN corpus. To select significant terms from the documents, we cluster the training documents using the EM algorithm [5] and extract terms that are found frequently in only one cluster.

To find the most appropriate values for the three parameters of the presented method we examined a range of possible values and extracted the best models based on their performance on the training part of each corpus separately. More specifically, for each corpus, we extracted the combination of parameter values that optimizes AUC in the training corpus examining the following range of values: feature type ∈ {word unigrams, char 3-grams, char 4-grams, char 5-grams}, $k \in \{20, 50, 70, 100, 150, \ldots, 500\}$, and similarity ∈ {cosine, weighted cosine, Euclidean, weighted Euclidean, Minmax, weighted Minmax} as defined in formulas 2, 3, and 4. The feature set contains all terms (word or character n-grams) occurring at least 5 times in the training corpus. Note that we only use information from the training part of the corpora to tune our approach examining each subset of problems of the training set belonging to a certain corpus separately and then we apply the tuned models to the corresponding test corpus. Therefore the presented evaluation results in terms of AUC of our method on the test corpus are directly comparable to the ones achieved by PAN participants[29, 30].

4.3 Results

First, we compare the performance of topic modeling approaches with the top-ranked submissions in PAN-2014 and PAN-2015 shared tasks. For each topic modeling approach (either LSI or LDA) we examine three variations: LSI_{train} and LDA_{train} refer to topic models extracted from the training corpus documents and their parameters also were optimized using the training corpus. In addition, LSI_{ext} and LDA_{ext} refer to topic models extracted by an enriched corpus where external documents were added to the training documents. Again, their parameters were optimized based on the training corpus. Last, $LSI_{ext}(200)$ and $LDA_{ext}(200)$ are topic models similar to the last ones but with a fixed number of topics (200). The performance of all these verification models on the test corpora is compared with the PAN-2014 and PAN-2015 winning submissions and with the work of Hernandez et al., [9] as well, as can be seen in Tables 1 and 2, respectively. Among these, the approaches of Moreau et al. [20] and Pacheco et al. [21] also use LDA features in combination with other types of features. In contrast, the method of Hernandez et al., [9] based exclusively on LDA features. The presented approach is entirely concerned with topic modeling features. In addition, the performance of another *intrinsic profile-based* approach as described by Potha and Stamatatos [22] is used as a baseline. This method has also been tuned for each PAN corpus separately.

In PAN-2014 corpora, the proposed verification methods based on LSI are particularly effective and outperform (in average scores) all top-ranked PAN-2014 methods and simultaneously the approach of Hernandez et al., [9]. The only exception is the PAN-2014-DR corpus which includes a minimal number of known documents and relatively short texts. Note that $avg(|D_{known}|)$ is the average number of known documents in a corpus. The size of D_{known} seems to be an important factor that influences the performance of the proposed methods. Given that our models follow the *profile-based paradigm*, concatenating all available known documents, it seems reasonable that their performance is improved when the number of known documents increases (like in the cases of PAN-14-EE, PAN-14-GR, and PAN-14-SP). However, it should be noticed that the method of Potha and Stamatatos [22], that is also an *intrinsic profile-based* approach, does not achieve notably better results when the size of D_{known} increases. This indicates that LSI is better able to handle long documents.

In general, LSI models are better than the corresponding LDA models (with the exception of PAN-14-EN). As concerns the performance of different variations of LSI models, it seems that the use of external documents helps the approach to become more stable. Moreover, the optimization of the number of topics based on the training set is helpful. However, relatively good results can also be obtained by using a fixed number of topics.

With respect to the results on PAN-2015 corpora, there is a more confused picture. In some cases (PAN-15-EN and PAN-15-SP) LSI models are very effective, competitive and even outperforming PAN-2015 winners, while in other cases (PAN-15-DU and PAN-15-GR) the performance of LSI models is not so competitive. Recall that in PAN-2015 corpora the texts within a verification instance may not belong to the same thematic area or genre. It seems therefore that these very challenging conditions significantly affect the performance of LSI models, especially in the cross-genre case of PAN-15-DU. On the other hand, the important improvement in the PAN-15-SP corpus that includes the highest size of D_{known} verifies the previous outcome that the proposed method is very effective when several known documents are available. Taking into consideration the corresponding results of the method of Hernandez et al., we

Table 1: AUC scores of the proposed and baseline models on the PAN-2014 corpora.

	EN	EE	DE	DR	GR	SP	avg
$avg(\lvert D_{known}\rvert)$	1.0	2.6	1.8	1.0	2.9	5.0	
Khonji & Iraqi (2014)	0.750	0.590	0.913	**0.736**	0.889	0.898	0.796
Frery et al.(2014)	0.612	0.723	0.906	0.601	0.679	0.774	0.716
Castillo et al.(2014)	0.628	0.549	0.861	0.669	0.686	0.734	0.688
Hernandez, et al. (2017)	0.644	0.780	0.855	0.577	0.686	0.576	0.686
Potha & Stamatatos (2014)	0.664	0.553	0.918	0.712	0.659	0.737	0.707
LDA_{train}	0.753	0.425	0.924	0.560	0.867	0.758	0.714
LDA_{ext}	0.752	0.620	0.820	0.451	0.794	0.628	0.678
$LDA_{ext}(200)$	**0.780**	0.552	0.934	0.479	0.797	0.630	0.695
LSI_{train}	0.696	0.748	0.960	0.686	**0.922**	0.890	0.817
LSI_{ext}	0.761	**0.781**	**0.982**	0.646	0.919	0.902	**0.832**
$LSI_{ext}(200)$	0.761	0.680	0.960	0.660	0.904	**0.906**	0.812

Table 2: AUC scores of the proposed and baseline models on PAN-2015 corpora.

	EN	DU	GR	SP	avg
$avg(\lvert D_{known}\rvert)$	1.0	1.76	2.93	4.0	
Bagnall (2015)	0.811	0.700	0.882	0.886	**0.820**
Moreau et al. (2015)	0.709	**0.825**	**0.887**	0.853	0.819
Pacheco et al.(2015)	0.763	0.822	0.773	0.908	0.817
Hernandez, et al. (2017)	**0.853**	0.751	0.709	0.783	0.774
Potha & Stamatatos (2014)	0.754	0.632	0.682	0.726	0.699
LDA_{train}	0.788	0.372	0.717	0.828	0.676
LDA_{ext}	0.681	0.593	0.770	0.810	0.714
$LDA_{ext}(200)$	0.624	0.538	0.760	0.816	0.685
LSI_{train}	0.811	0.639	0.836	0.750	0.759
LSI_{ext}	0.802	0.589	0.851	0.919	0.790
$LSI_{ext}(200)$	0.764	0.572	0.859	**0.946**	0.785

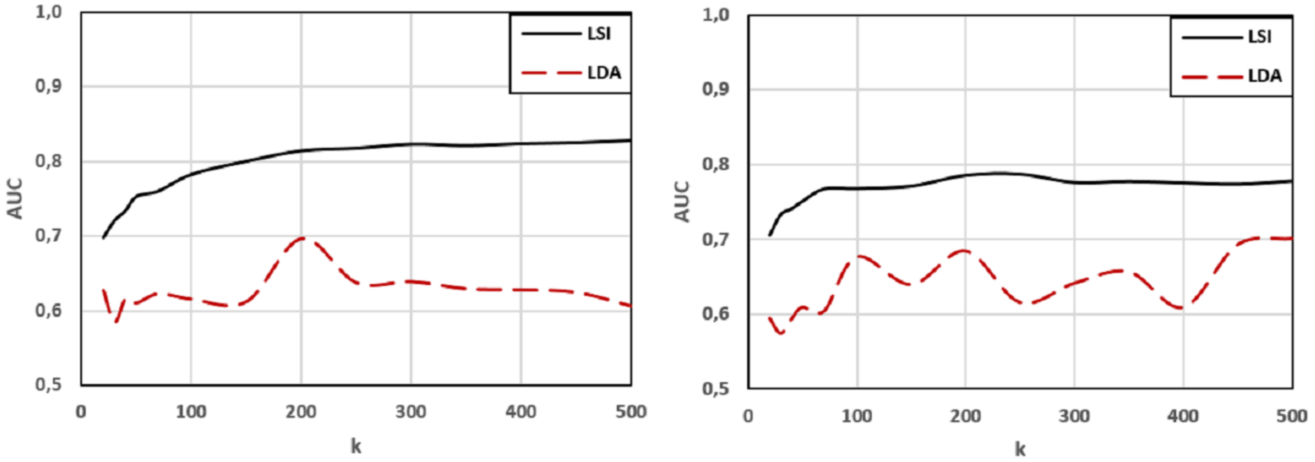

Figure 1: Average performance of LSI_{ext} and LDA_{ext} models on PAN-2014 (left) and PAN-2015 (right) corpora.

notably reinforce the aforementioned outcome as our proposed verification models are also very effective and clearly better options in the PAN-15-GR and PAN-15-SP corpus where the number of D_{known} is particularly enhanced. On the other hand, LSI models are proved to be not competitive enough in the case of PAN-15-EN where only one known document is available per problem and continue performing poorly results in PAN-15-DU.

In contrast, the other tested *profile-based* method [22] does not seem to be positively influenced when multiple known documents are available. Again, LSI models outperform LDA models, with a wide margin in most of the cases. Similar to PAN-2014 results, the

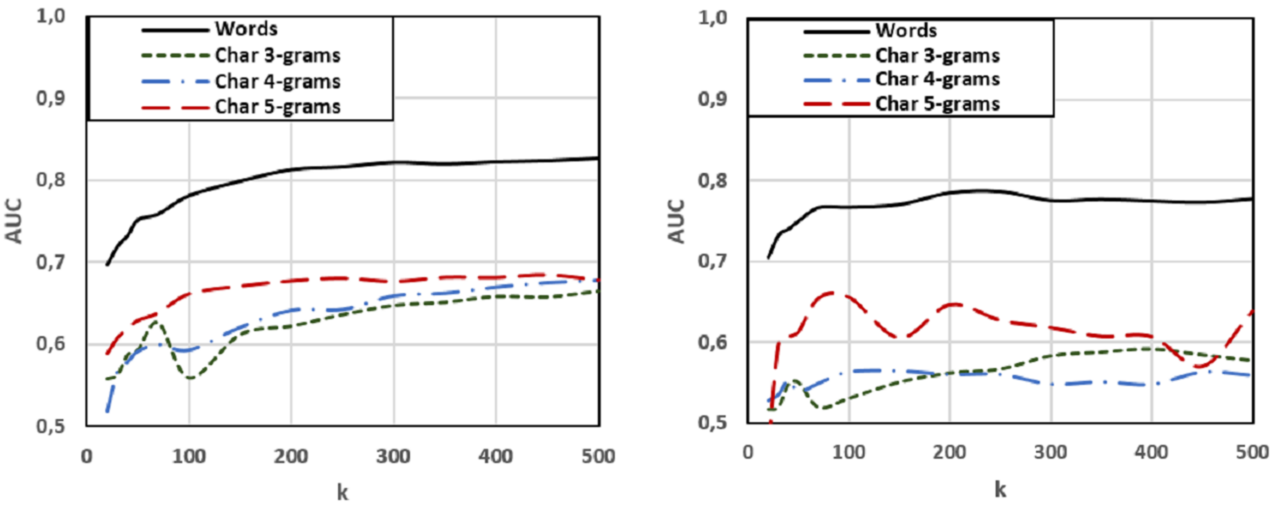

Figure 2: Average performance (AUC) of LSI_{ext} model with cosine similarity on PAN-2014 (left) and PAN-2015 (right) corpora for different feature types.

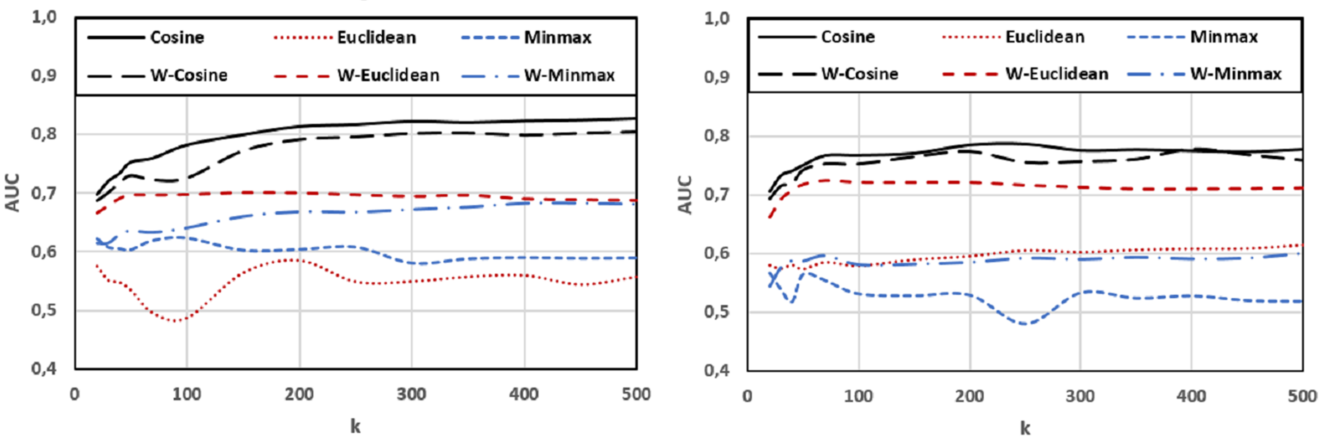

Figure 3: Average performance (AUC) of LSI_{ext} model with word unigram features on PAN-2014 (left) and PAN-2015 (right) corpora for different similarity measures.

use of additional external documents to extract the topic models seems to be helpful.

In order to demonstrate the usefulness of LSI and LDA models in the proposed *profile-based* author verification approach, Figure 1 shows the average performance (AUC) of LSI and LDA models in PAN-2014 and PAN-2015 corpora for a varying number of topics. As can be clearly seen, LSI models are more effective in the whole range of examined topics in both PAN-2014 and PAN-2015 corpora. Moreover, the performance of LSI is more stable since it is not affected too much when at least 100 topics are extracted.

Next, we focus on the contribution of feature types when LSI models are used. Figure 2 shows the average performance of LSI_{ext} with cosine similarity for a varying number of topics on PAN-2014 and PAN-2015 corpora when different feature types are examined. It is clear that word unigram features are the best option and they provide a more stable performance curve. As concerns character *n*-grams, relatively long *n*-grams (5-grams) seem to be more effective than short *n*-grams (3-grams and 4-grams) that are commonly used in authorship analysis studies [28].

Finally, Figure 3 demonstrates the average performance of LSI_{ext} with word unigram features on PAN-2014 and PAN-2015 corpora when different similarity measures are used. In both cases, cosine similarity is far better than Euclidean and Minmax similarity. Although, the weighted version of cosine similarity is competitive enough, the simple (unweighed) version consistently outperforms its performance. On the other hand, for Euclidean and Minmax, their weighted version is much better than the unweighed one. The weighted Euclidean similarity is also quite stable when more than 50 topics are used.

5 CONCLUSION

In this paper, we focused on the use of topic modeling in the author verification task. Based on an intrinsic and profile-based approach to author verification, we demonstrated that LSI is better than LDA because it is both more effective and more stable when the number of topics varies. As demonstrated in the performed experiments, LSI models work better with word unigram features and cosine similarity. It is also useful to extract the topic models from a large corpus with similar properties with the texts under investigation.

Our proposed approach is easy-to-follow and language independent. Moreover, it has a clear advantage over alternative methods when multiple documents of known authorship are available per verification instance. This can be partially explained by the fact that the *profile-based paradigm* is followed. In addition, LSI models seem to be more effective with longer texts. The performance of the proposed method seems to be affected in cross-domain conditions, especially in the case of cross-genre verification. This indicates that a more sophisticated approach should be designed for these challenging cases, perhaps using genre-specific language models. Another future work dimension could be to explore the use of the proposed method in the framework of an *extrinsic* verification method.

REFERENCES

[1] Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20, 5 (2005), 67–75.

[2] Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891* (2015).

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41, 6 (1990), 391.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39, 1 (1977), 1–38.

[6] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[7] Jordan Fréry, Christine Largeron, and Mihaela Juganaru-Mathieu. 2014. Ujm at clef in author identification. *Proceedings CLEF-2014, Working Notes* (2014), 1042–1048.

[8] Oren Halvani, Christian Winter, and Anika Pflug. 2016. Authorship verification for different languages, genres and topics. *Digital Investigation* 16 (2016), S33 – S43.

[9] Ángel Hernández-Castañeda and Hiram Calvo. 2017. Author Verification Using a Semantic Space Model. *Computación y Sistemas* 21, 2 (2017).

[10] Magdalena Jankowska, Evangelos Milios, and Vlado Keselj. 2014. Author verification using common n-gram profiles of text documents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 387–397.

[11] Patrick Juola. 2008. Authorship Attribution. *Foundations and Trends in Information Retrieval* 1 (2008), 234–334. Issue 3.

[12] Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. In *Working Notes for CLEF 2013 Conference*.

[13] Mahmoud Khonji and Youssef Iraqi. 2014. A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF). In *CLEF 2014 Labs and Workshops, Notebook Papers*. CLEF and CEUR-WS.org.

[14] April Kontostathis and William M Pottenger. 2006. A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management* 42, 1 (2006), 56–73.

[15] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation* 45, 1 (2011), 83–94.

[16] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Yaron Winter. 2012. The ???Fundamental Problem??? of Authorship Attribution. *English Studies* 93, 3 (2012), 284–291.

[17] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research* 8 (2007), 1261–1276.

[18] Moshe Koppel and Shachar Seidman. 2013. Automatically Identifying Pseudepigraphic Texts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1449–1454.

[19] Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the American Society for Information Science and Technology* 65, 1 (2014), 178–187.

[20] Erwan Moreau, Arun Jayapal, Gerard Lynch, and Carl Vogel. 2015. Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners-Notebook for PAN at CLEF 2015. In *CLEF 2015-Conference and Labs of the Evaluation forum*. CEUR.

[21] María Leonor Pacheco, Kelwin Fernandes, and Aldo Porco. 2015. Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification.. In *CLEF (Working Notes)*.

[22] Nektaria Potha and Efstathios Stamatatos. 2014. A Profile-Based Method for Authorship Verification. In *Artificial Intelligence: Methods and Applications - Proceedings of the 8th Hellenic Conference on AI, SETN*. 313–326.

[23] Nektaria Potha and Efstathios Stamatatos. 2017. An Improved Impostors Method for Authorship Verification. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 138–144.

[24] Conrad Sanderson and Simon Guenter. 2006. Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. In *Proceedings of the International Conference on Empirical Methods in Natural Language Engineering*. 482–491.

[25] Jacques Savoy. 2013. Authorship attribution based on a probabilistic topic model. *Information Processing and Management* 49, 1 (2013), 341–354.

[26] Shachar Seidman. 2013. Authorship Verification Using the Impostors Method. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, Pamela Forner, Roberto Navigli, and Dan Tufis (Eds.).

[27] Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics* 40, 2 (2014), 269–310.

[28] Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60 (2009), 538–556. Issue 3.

[29] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2014. Overview of the Author Identification Task at PAN 2014.. In *CLEF (Working Notes)*. 877–897.

[30] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. Overview of the Author Identification Task at PAN 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*.

[31] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics* 26, 4 (2000), 471–495.

[32] Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the American Society for Information Science and Technology* 67, 1 (2016), 239–242.

[33] Jianwen Sun, Zongkai Yang, Sanya Liu, and Pei Wang. 2012. Applying Stylometric Analysis Techniques to Counter Anonymity in Cyberspace. *JNW* 7, 2 (2012), 259–266.