Improving Author Verification Based on Topic Modeling

Nektaria Potha and Efstathios Stamatatos

Department of Information and Communication Systems Engineering,

University of the Aegean

Karlovassi, Samos, 83200, Greece

{nekpotha,stamatatos}@aegean.gr

Abstract

Authorship analysis attempts to reveal information about authors of digital documents enabling applications in digital humanities, text forensics, and cyber-security. Author verification is a fundamental task where given a set of texts written by a certain author we should decide whether another text is also by that author. In this paper, we systematically study the usefulness of topic modeling in author verification. We examine several author verification methods that cover main paradigms, namely intrinsic (attempt to solve a one-class classification task) and extrinsic (attempt to solve a binary classification task) methods as well as profile-based (all documents of known authorship are treated cumulatively) and instance-based (each document of known authorship is treated separately) approaches combined with well-known topic modeling methods such as Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA). We use benchmark datasets and demonstrate that LDA is better combined with extrinsic methods while the most effective intrinsic method is based on LSI. Moreover, topic modeling seems to be particularly effective for profile-based approaches and the performance is enhanced when latent topics are extracted by an enriched set of documents. The comparison to state-of-the-art methods demonstrates the great potential of the approaches presented in this study. It is also demonstrated that even in case genre-agnostic external documents are used, the proposed extrinsic models are very competitive.

Keywords: Author verification, Authorship analysis, Stylometry, Text categorization, Text mining

Improving Author Verification Based on Topic Modeling

Introduction

Nowadays, there are massive amounts of texts in digital form in digital libraries, online journalism and social networks. For example, it is estimated that half billion tweets are sent per day. This underlines the need for handling this information efficiently. Automated text categorization plays a crucial role in this process (Aggarwal & Zhai, 2012) where topic, sentiment and style of documents can be used as discriminating factors. In particular, style of documents can be used to infer their genre or reveal information about their authors (Stamatatos, 2009). Authorship analysis, dealing with the personal style of authors, is a very active research area (Neal et al., 2017; Rocha et al., 2017; Seroussi, Zukerman, & Bohnert, 2014; Stamatatos, 2018).

Among authorship analysis tasks, author verification has an eminent position. Formally, given a set of sample documents of known authorship (D_{known}), all by the same author, and another document of unknown authorship (d_u), the question is whether or not the latter is also by that candidate author (Juola & Stamatatos, 2013; Koppel & Winter, 2014). An author verification method attempts to estimate the probability $p(d_u|D_{known})$ indicating how likely it is for d_u to be written by the author of D_{known}. Actually, author verification is a special instance of open-set authorship attribution where the set of candidate authors is singleton. However, any authorship attribution case, either closed-set or open-set, can be decomposed into a series of author verification problems (Koppel, Schler, Argamon, & Winter, 2012). This demonstrates the fundamental role of this task.

In the relevant literature, there are several examples of applications associated with author verification technology. In digital humanities, author verification methods have been used to reveal the authorship of disputed documents of great literary or historical importance (Kestemont, Stover, Koppel, Karsdorp, & Daelemans, 2016; Stover, Winter, Koppel, & Kestemont, 2016; Tuccinardi, 2017). In cyber-security, it has been used to detect compromised accounts in social networks (Barbon, Igawa, & Bogaz Zarpelão, 2017), perform spearphishing filtering (Duman, Kalkan-Cakmakci, Egele, Robertson, & Kirda, 2016), and continuous authentication of users (Brocardo, Traore, Woungang, & Obaidat, 2017). In

deception detection, author verification has been used to identify fake reviews by providing evidence that reviews published under different aliases are actually written by the same author (Layton, Watters, & Ureche, 2013). This technology can also be used to enhance recommender systems (Vaz, Martins de Matos, & Martins, 2012), opinion mining (Panicheva, Cardiff, & Rosso, 2010), personalized spam email detection (Shams & Mercer, 2016), and spun content detection (Shahid et al., 2017).

Recently, a variety of author verification methods have been developed (Bagnall, 2015; Ding, Fung, Iqbal, & Cheung, 2018; Jankowska, Milios, & Keselj, 2014; Halvani, Winter, & Pflug, 2016; Fréry, Largeron, & Juganaru-Mathieu, 2014; Kocher & Savoy, 2017; Koppel & Winter, 2014). A series of PAN[1] shared tasks have also contributed to increase the interest of research community about this task (Juola & Stamatatos, 2013; Stamatatos et al., 2014, 2015). In general, there are specific paradigms that most author verification methods follow. Depending on the way they handle the available samples in D_{known}, verification methods fall under one of the following paradigms (Stamatatos, 2009), depicted in Figure 1:

1. *Instance-based paradigm*: Each of the available samples in D_{known} is represented separately and is treated as a distinct instance of the author's style (Jankowska et al., 2014; Khonji & Iraqi, 2014; Seidman, 2013). The approaches in this paradigm are document-centric and attempt to explore and take advantage of differences between documents by the same author.

2. *Profile-based paradigm*: All available samples in D_{known} are first concatenated in one big document and a single representation vector is extracted (author's profile) (Ding et al., 2018; Halvani et al., 2016; Kocher & Savoy, 2017; Potha & Stamatatos, 2014). The methods in this paradigm follow an author-centric approach where the differences between documents by the same author are disregarded.

From another point of view, author verification methods can be distinguished according to the set of documents they analyze. There are two basic paradigms depicted in Figure 2:

---

[1] This acronym originates from a SIGIR-2007 workshop entitled *Plagiarism analysis, Author identification and Near-duplicate detection* (Stein, Koppel, & Stamatatos, 2007)

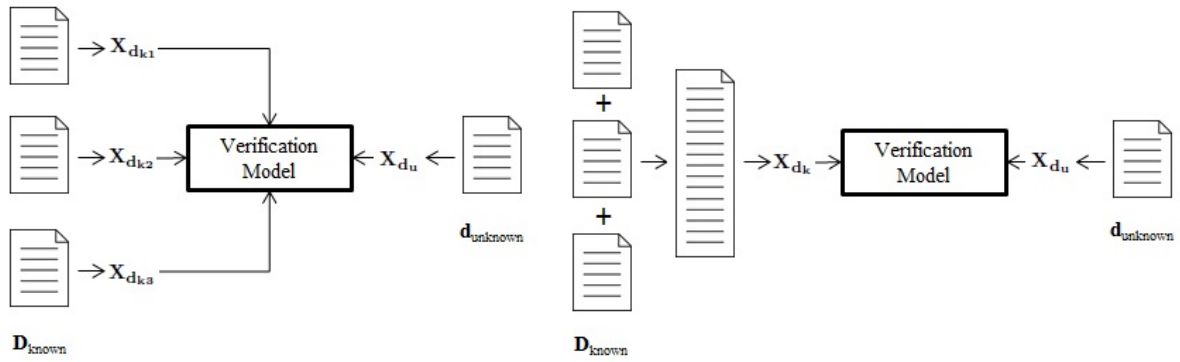

Figure 1. Instance-based (left) vs. profile-based (right) verification methods.

1. *Intrinsic verification methods*: they only analyze samples in D_{known} and d_u. Essentially, they consider author verification as a one-class classification task and attempt to estimate whether d_u is similar enough to D_{known} (Halvani et al., 2016; Jankowska et al., 2014; Potha & Stamatatos, 2014). Intrinsic models are usually robust since they do not depend on external resources.

2. *Extrinsic verification methods*: in addition to D_{known} and d_u they also consider a set of external documents by other authors $D_{external}$. They attempt to transform author verification to a binary or multi-class classification task by estimating whether the similarity between d_u and D_{known} is higher than the similarity between d_u and members of $D_{external}$ (Koppel & Winter, 2014; Potha & Stamatatos, 2017; Seidman, 2013). Extrinsic models can be very effective when $D_{external}$ is carefully selected.

A key point in author verification (and authorship analysis in general) is to adequately represent the personal style of authors. To this end, stylometric measures are extracted from documents. Several kinds of such measures exist and they may be computationally simplistic (e.g., word and character n-gram frequencies) or more sophisticated (e.g., demanding the application of natural language processing tools, like part-of-speech (POS) taggers and full syntactic parsers) (Stamatatos, 2009). Usually, several hundreds or thousands of such measures are used and the resulting representation is quite sparse.

Topic modeling is a well-known approach to reduce the dimensionality and provide a non-sparse representation of documents (Blei, Ng, & Jordan, 2003; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Topic modeling methods attempt to discover hidden

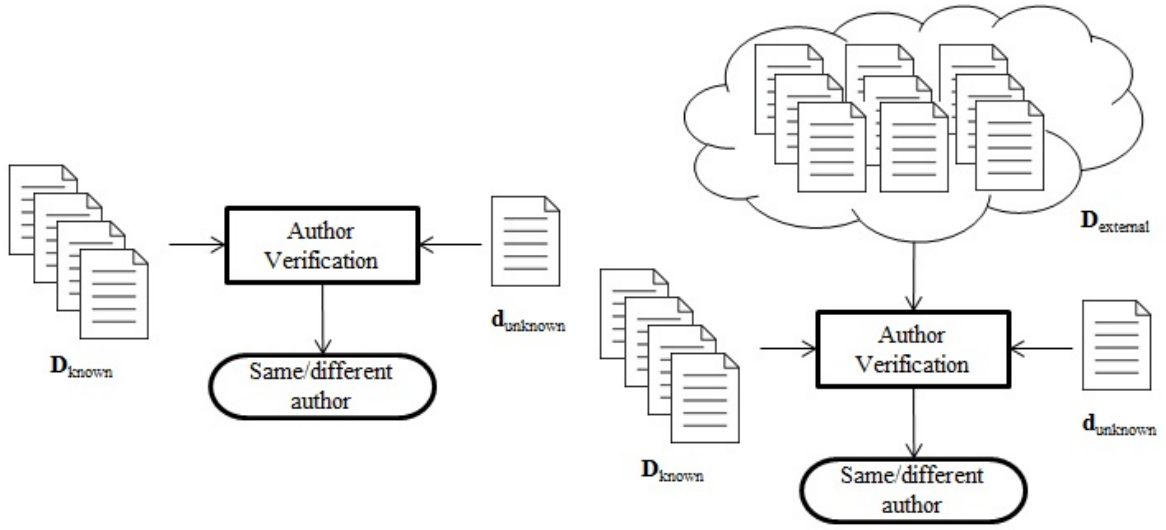

Figure 2. Intrinsic (left) vs. extrinsic (right) verification methods.

structures in the texts, that can be viewed as latent semantic structures (or latent topics). Although the first applications of such methods were related to thematic text classification, it became clear that the extracted structures can also capture stylistic information. For example, Seroussi et al. (2014) report that some latent topics extracted from movie reviews and message board posts are related to the use of colloquial words while others are related to more formal words. Topic modeling methods have already been applied to authorship analysis tasks like authorship attribution (Savoy, 2013; Seroussi et al., 2014) and author profiling (López-Monroy, Montes-y-Gómez, Escalante, Villaseñor-Pineda, & Stamatatos, 2015). In author verification, topic modeling has only been applied occasionally (Hernández & Calvo, 2017; Pacheco, Fernandes, & Porco, 2015; Potha & Stamatatos, 2018) so far. Such works do not systematically study the effect of different topic modeling techniques to various verification paradigms. As a result, it is not yet clear if all author verification paradigms are improved by exploiting topic modeling. In addition, it is not clear what specific topic modeling technique is the most appropriate for each individual author verification paradigm.

The main contributions of this study are listed below:

- A systematic study is presented that examines the usefulness of topic modeling when combined with author verification methods covering both instance-based and profile-based paradigms as well as both intrinsic and extrinsic paradigms. We focus on

the two most well-known topic modeling approaches, namely Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) and examine their suitability for each author verification paradigm.

- It is demonstrated that topic modeling can considerably increase the effectiveness of author verification methods when an appropriate topic modeling technique is selected and is adequately fine-tuned.

- We examine the effect of using either a limited set of documents or an enriched document collection to extract latent topics and it is demonstrated that the latter assists author verification methods to further increase their effectiveness.

- We report experimental results on benchmark datasets developed during the relevant PAN-2014 and PAN-2015 shared tasks in author verification that are directly compared with state-of-the-art methods under the same settings. The performance of the methods presented in this study is quite competitive to the best results reported so far for these datasets demonstrating that topic modeling can be an efficient and effective alternative to more sophisticated methods (e.g, based on representation learning, distributed document representation, or neural network language models (Bagnall, 2015; Ding et al., 2018)) for the author verification task.

- We examine the effect of genre of external documents when extrinsic author verification methods are combined with topic modeling techniques. It is demonstrated that verification models based on genre-agnostic external documents are very competitive but they are outperformed by models using external documents of the same genre with that of the questioned documents.

The rest of this paper is organized as follows. In section 2 a review of previous author verification studies is presented. Section 3 describes the set of examined author verification methods as well as the used topic modeling techniques. Then, Section 4 focuses on the performed experiments while Section 5 discusses the main conclusions drawn from this study and possible future work directions.

Previous Work

An early study on author verification is presented by Stamatatos, Fakotakis, and Kokkinakis (2000). They use stylistic measures extracted from a natural language processing tool to perform text genre detection, closed-set authorship attribution, and author verification. A seminal approach, known as *unmasking* method was introduced by Koppel, Schler, and Bonchek-Dokow (2007). This iterative method attempts to distinguish between two documents and it measures how quickly cross-validation accuracy degrades when a specific amount of significant discriminative features are removed in each iteration. If the accuracy drops fast, then the two documents are not by the same author. This method provides exceptional results when handling long documents (e.g., novels). However, when dealing only with short documents or when documents belong to different genres, it practically fails (Kestemont, Luyckx, Daelemans, & Crombez, 2012; Sanderson & Guenter, 2006).

In recent years, the popularity of author verification task has noticeably grown up, mainly due to the focus of relevant PAN shared tasks organized for three consecutive years (2013-2015) (Juola & Stamatatos, 2013; Stamatatos et al., 2014, 2015). A series of benchmark corpora were released that period covering several natural languages, genres, and degrees of difficulty.

As already explained, intrinsic verification methods consider exclusively the available documents in a verification case and they attempt to solve a one-class classification task where only positive examples (D_{known}) are given (Halvani et al., 2016; Jankowska et al., 2014; Potha & Stamatatos, 2014). These methods calculate the similarity of d_u to D_{known} and then they decide if this similarity is significant. The latter can be determined by using information from other verification cases (Jankowska et al., 2014; Potha & Stamatatos, 2014). A recent study achieves to determine all necessary thresholds without using any additional resources (Halvani, Graner, & Vogel, 2018). In general, intrinsic methods are very efficient and robust methods with few parameters to be set and can be easily applied to any corpus.

On the contrary, extrinsic methods take advantage of additional documents written by other authors collected from external resources to transform author verification to a binary classification task (Bagnall, 2015; Koppel & Winter, 2014; Pacheco et al., 2015). Namely,

D_{known} represents the positive class, $D_{external}$ represents the negative class, and d_u is assigned to one of them. The most influential method of this category is called *Impostors* and was introduced by Koppel and Winter (2014). Several modifications of this method have been reported to achieve remarkable results (Khonji & Iraqi, 2014; Potha & Stamatatos, 2017; Seidman, 2013). It is remarkable that the top-performing approaches in all three PAN shared tasks in author verification follow this paradigm (Bagnall, 2015; Khonji & Iraqi, 2014; Seidman, 2013). However, recent studies show that intrinsic methods can be equally competitive (Ding et al., 2018; Halvani, Winter, & Graner, 2017; Potha & Stamatatos, 2018). Extrinsic methods are heavily influenced by the suitability of documents in $D_{external}$ for a given verification case. It is questionable whether the selected external documents are representative of the negative class (Stein, Lipka, & z. Eissen, 2008). In addition, when the external documents belong to a different genre with respect to D_{known} and d_u, the performance of extrinsic methods is negatively affected (Koppel & Winter, 2014).

It should be noted that there is another approach in author verification studies that also attempts to solve a binary classification task, however, in a different way in comparison to external verification methods. In more detail, if a set of training instances is available, each instance is a tuple $(d_u, D_{known}, class)$, then it is possible to build a classifier that distinguishes between two classes: positive (same author) and negative (different author). A supervised learning algorithm can be used to learn a general verification model (Fréry et al., 2014; Hürlimann, Weck, van den Berg, Šuster, & Nissim, 2015; Pacheco et al., 2015). These *eager* verification approaches (Stamatatos, 2016) can follow either the intrinsic paradigm (each verification instance is represented using only information from d_u and D_{known}) (Hürlimann et al., 2015) or the extrinsic paradigm (each instance is represented using information from d_u, D_{known}, and $D_{external}$) (Pacheco et al., 2015). The effectiveness of such methods heavily depend on the representativeness of the training set of instances (Stamatatos et al., 2015).

From a different perspective, author verification methods can be distinguished between instance-based and profile-based ones (see Fig. 1). The former attempts to discover and exploit differences in the set of documents of known authorship (Jankowska et al., 2014; Khonji & Iraqi, 2014; Seidman, 2013). The majority of PAN submissions follow this

paradigm (Stamatatos, 2016). On the other hand, profile-based approaches handle all available known texts cumulatively (Kocher & Savoy, 2017; Pacheco et al., 2015; Potha & Stamatatos, 2014). Profile-based methods are better able to handle cases with limited text length since they concatenate all available documents. This paradigm is more popular in recent studies (Ding et al., 2018; Halvani et al., 2018; Potha & Stamatatos, 2018). There are also some attempts to combine instance-based and profile-based paradigms (Bagnall, 2015; Sari & Stevenson, 2015).

It is also important to note that ensembles of author verification methods seem to be a very effective solution. The organizers of PAN shared tasks report the performance of a simple meta-model averaging the output of all submitted approaches and in many cases it outperforms any individual methods (Juola & Stamatatos, 2013; Stamatatos et al., 2014, 2015). Similarly, another heterogeneous ensemble that combines five distinct author verification approaches achieved encouraging results (Moreau, Jayapal, Lynch, & Vogel, 2015).

With respect to text representation, author verification methods are usually based on simple measures like character-level features (i.e., letter frequencies, punctuation mark frequencies, character n-grams, etc.) and lexical features (i.e., word frequencies, word n-grams, function word frequencies, etc.) (Stamatatos et al., 2014, 2015) Such features are easily extracted from documents and are language-independent essentially. A particularly effective approach is based on a character-level recurrent neural network language model (Bagnall, 2015). This method achieved the best overall results in PAN-2015 shared task but it seems to be confused when documents within a verification case belong to different genres (Stamatatos et al., 2015).

The use of more sophisticated features related to syntactic analysis or semantic analysis of documents is very limited so far while the reported results of methods based on such features are not particularly encouraging (Stamatatos et al., 2015). Usually, the most sophisticated features found in author verification studies are related to POS tag frequencies (Khonji & Iraqi, 2014; Pacheco et al., 2015). However, they heavily depend on the accuracy and suitability of POS tagger used to analyze the documents.

Topic modeling has been used only occasionally so far in author verification studies. In some cases, LDA features are used as a complement to other basic feature types (Moreau et

al., 2015; Pacheco et al., 2015). A recent study is exclusively based on LDA features and attempts to discover patterns of latent topic usage in known and unknown documents of an author verification case (Hernández & Calvo, 2017). Moreover, in another recent work, LSI was found to be more effective than LDA for an intrinsic and profile-based method (Potha & Stamatatos, 2018). In this paper we extend the latter and examine the usefulness of both LSI and LDA for a variety of author verification paradigms.

A remarkable recent approach is based on representation learning. Ding et al. (2018) examined several modalities, namely character, lexical, topical, and syntactic (POS tags) and a neural network that can jointly learn multiple modalities. The most effective model was based on lexical and topical modalities that reflect both the global topic of the document and the personal bias of the author in choosing specific words given that topic. This approach has a relatively large number of hyper-parameters to be set. The experimental results on PAN-2014 benchmark datasets demonstrated that this sophisticated method can outperform existing methods as well as baselines based on topic modeling techniques (following an intrinsic and profile-based paradigm with a pre-fixed number of latent topics) as well as other distributed word and document representations (word2vec, doc2vec) (Ding et al., 2018). In this paper, we show that similar or even higher performance can be achieved by appropriately combining fine-tuned topic modeling techniques and author verification paradigms.

Author Verification Methods

Initially, in all author verification methods we consider in this study, all documents are represented as vectors of normalized frequencies of word unigrams. Let $\vec{X}_d$ be the representation vector of a document d in this high dimensional space. Then, a topic modeling method estimates a latent semantic space and derives $\hat{\vec{X}}_d$, a non-sparse representation of d in a new space of considerably reduced dimensionality. In addition, when a similarity between two documents d_1 and d_2 is required, the cosine similarity of the reduced vectors is used:

$$similarity(d_1, d_2) = cosine(\hat{\vec{X}}_{d_1}, \hat{\vec{X}}_{d_2}) \quad (1)$$

In the following we describe the author verification methods used in this study covering

both intrinsic and extrinsic methods as well as instance-based and profile-based paradigms.

Intrinsic Verification Approaches

Instance-based Method. The intrinsic instance-based verification method is inspired from Jankowska et al. (2014), a very robust method that was used as baseline in PAN-2014 and PAN-2015 shared tasks (Stamatatos et al., 2014, 2015). Initially, all pairwise similarities in D_{known} are estimated and the minimum similarity between each known document d_i and the rest of known documents $SimMin(d_i, D_{known})$ is extracted. Then, the similarity between each d_i and d_u is estimated and divided by $SimMin(d_i, D_{known})$. This ratio indicates whether d_i is more similar to the unknown document with respect to the rest of known documents. Finally, the average of such ratios is calculated over all available known documents:

$$verificationScore(d_u, D_{known}) = \frac{\sum_{i=1}^{|D_{known}|} (\frac{similarity(d_i, d_u)}{SimMin(d_i, D_{known})})}{|D_{known}|} \quad (2)$$

Since the produced verification score can take values larger than 1, it can be calibrated based on a set of training instances I_{train} (verification cases of similar properties, e.g., language and genre):

$$p(d_u|D_{known}) \sim calibrate(verificationScore(d_u, D_{known}), I_{train}) \quad (3)$$

It should be noted that when $|D_{known}| = 1$ then the only available known document is split in two parts of equal length so that $SimMin$ can be computed. In the rest of this paper, we call this method IIB (intrinsic and instance-based method).

Profile-based Method. Following the profile-based paradigm we adopt a simple but effective method described by Potha and Stamatatos (2018). First we concatenate all available samples in D_{known} in a single document d_k. Then, d_k is compared with d_u and their similarity indicates the likelihood of both being written by the same author:

$$p(d_u|D_{known}) \sim similarity(d_k, d_u) \quad (4)$$

This approach can be easily applied to any verification problem and it is not so much

affected by limited text length since the concatenation of known documents produces a relatively long document. In the rest of this paper, we call this method IPB (intrinsic and profile-based method). Note also that both intrinsic methods used in this study have no internal hyper-parameters that should be fixed apart from the number of latent topics that is actually related with the topic modeling technique with which it is combined. Therefore, their application to any benchmark corpus is straightforward.

Extrinsic Approaches

Instance-based Method. All extrinsic methods need a set of external documents $D_{external}$ by authors other than the author of D_{known}, ideally carefully selected to be on the same topic and belong to the same genre with documents of D_{known}. The most well-known representative of extrinsic verification methods is the *Impostors* method introduced by Koppel and Winter (2014). This method is essentially a random subspace ensemble where in each repetition a random subset of external documents and a random number of features is selected. Then, the similarity between d_u and each $d \in D_{known}$ as well as the similarity between d_u and each $d \in D_{external}$ (aka impostor) is calculated. Then, the percentage of times the unknown document was found more similar to the known documents rather than the impostors is counted. The higher this percentage, the more likely the unknown document to be written by the author of D_{known}.

Since the original Impostors method (Koppel & Winter, 2014) can only handle cases where $|D_{known}| = 1$, there are variants that can be applied to any number of known documents. One of them is called General Impostors (GI) (Seidman, 2013) and simply applies the Impostor method for each known document separately and finally aggregates the results. In this study, we use another variant called ranking-based Impostors (Potha & Stamatatos, 2017) that also takes into account the rank of $similarity(d_i, d_u)$ for each $d_i \in D_{known}$ with respect to the similarities between d_u and each $d \in D_{external}$. If $similarity(d_i, d_u)$ is near the top then this also contributes significantly in the calculation of the verification score which corresponds to the final output of the algorithm:

$$p(d_u|D_{known}) \sim VerificationScore(d_u, D_{known}, D_{external}) \quad (5)$$

This variant of Impostors is illustrated in algorithm 1. Note that the $relevance$ function is used to rank impostors by their similarity to known documents while the notation $relevance(:)$ denotes all available relevance values. In the rest of this paper, this method is called EIB (extrinsic and instance-based method). Note that it accepts exactly the same data and parameters in comparison to GI. In addition, all variants of Impostors are stochastic algorithms (since they make some random choices). Their hyper-parameters should be tuned based on a training corpus as we will explain later.

Profile-based Method. In this paper, we propose another variation of the Impostors method that follows the profile-based paradigm where all known documents are concatenated and the differences between them are disregarded. That way, there is only one long known document d_k and there is no need to aggregate results for separate known documents. On the other hand, d_k should be compared with appropriate external documents. Given that d_k is composed by several constituent documents, a fact affecting its thematic homogeneity, the impostors should also have similar characteristics. To achieve this, only the external documents with the highest cumulative similarity to the set of known documents are considered. That way, external documents of distinct style and topic with respect to the specific known documents are filtered out. In contrast to EIB, the selected external documents should be similar enough with all known documents. That is, if an external document is very similar with one known document but highly dissimilar with the rest of them, it is not likely to be selected. Thus, the selected documents are more likely to have stylistic rather than thematic similarities with the known documents.

Each impostor is then constructed by selecting one of these documents highly similar to all members of D_{known} randomly. Then, its nearest neighbors (using cosine similarity) are found so that we have exactly $|D_{known}|$ documents. The concatenation of these documents corresponds to an artificial impostor document with similar characteristics with the known text, namely it is composed by the same number of constituent documents which have certain stylistic similarities. Similar to EIB, the ranking information of the known document within

Data: D_{known}, $d_{unknown}$, $D_{external}$
Parameters: $repetitions$,$|Impostors_{problem}|$,$|Impostors_{repetition}|$,$rate$
Result: $VerificationScore$
for *each* $d_{known} \in D_{known}$ **do**
  for *each* $impostor \in D_{external}$ **do**
    $relevance(impostor) = similarity(impostor, d_{known})$;
  end
  Select $Impostors_{problem} \subset D_{external}$ with highest $relevance(:)$;
  Set $Score(d_{known}) = 0$;
  repeat $repetitions$ times
    Select $Impostors_{repetition} \subset Impostors_{problem}$ randomly;
    Select $rate\%$ of features randomly;
    for *each* $impostor \in Impostors_{repetition}$ **do**
      $Sim(impostor) = similarity(impostor, d_{unknown})$;
    end
    $Sim_{known} = similarity(d_{known}, d_{unknown})$;
    Rank $S = Sim(:) \cup Sim_{known}$ in decreasing order;
    $pos =$ position of Sim_{known} in S;
    $Score(d_{known}) = Score(d_{known}) + 1/(repetitions * pos)$;
  end;
end
$VerificationScore = aggregate(Score(:))$;

Algorithm 1: The extrinsic and instance-based verification method used in this study.

the set of impostors is considered in order to calculate the verification score. This profile-based Impostors approach is demonstrated in algorithm 2. Note that it accepts the same data and hyper-parameters with EIB. In the rest of this paper this method is called EPB (extrinsic and profile-based method).

Topic Models

Topic modeling is one of the most well-known techniques to decisively reduce dimensionality of document representation that can be easily applied to large volumes of data. Topic modeling maps the original high-dimensional and sparse feature space into a small set of new features that correspond to latent semantic structures (topics) in documents. Each document is then represented as a mixture of these topics. Several topic modeling algorithms have been applied in text categorization tasks for dimensionality reduction and providing a less sparse representations of texts. That way, the reduced space is less redundant, the resulting data are more compact and less noisy. In this study, we consider the two most widely used topic modeling techniques, LSI and LDA.

Data: D_{known}, $d_{unknown}$, $D_{external}$
Parameters: $repetitions$,$|Impostors_{problem}|$,$|Impostors_{repetition}|$,$rate$
Result: $VerificationScore$
Set $relevance(:) = 0$;
for *each* $d_{known} \in D_{known}$ **do**
  for *each* $impostor \in D_{external}$ **do**
    $relevance(impostor) = relevance(impostor) + similarity(impostor, d_{known})$;
  end
end
Select $Impostors_{problem} \subset D_{external}$ with highest $relevance(:)$;
Set $VerificationScore = 0$;
$knownText = concatenate(D_{known})$;
repeat $repetitions$ times
  Select $rate\%$ of features randomly;
  $Sim_{known} = similarity(knownText, d_{unknown})$;
  repeat $|Impostors_{repetition}|$ times
    Select $impostorSeed \in Impostors_{problem}$ randomly;
    Select $Neighbours \subset Impostors_{problem}$, the $|D_{known}| - 1$ most similar to $impostorSeed$;
    $newImpostor$ = concatenate($impostorSeed \cup Neighbours$);
    $Sim(newImpostor) = similarity(newImpostor, d_{unknown})$;
  end;
  Rank $S = Sim(:) \cup Sim_{known}$ in decreasing order;
  $pos =$ position of Sim_{known} in S;
  $VerificationScore = VerificationScore + 1/(repetitions * pos)$;
end;

Algorithm 2: The extrinsic and profile-based method used in this study.

1. LSI is a deterministic and algebraic topic modeling technique. It attempts to uncover the latent structures in a collection of documents by exploiting word co-occurrences in documents. In the approximation of a small dimensional space LSI is 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 (topics) 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 (Deerwester et al., 1990). Hence, LSI needs a parameter to be set, namely the number of topics (size of the reduced space).

2. LDA is a generative probabilistic topic modeling approach (Blei et al., 2003). 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 follows a sparse Dirichlet prior. LDA also needs the number of topics to be tuned and it is more popular than LSI in previous authorship verification studies (Moreau et al., 2015; Pacheco et al., 2015; Hernández & Calvo, 2017).

To appropriately tune the number of latent topics, a training corpus can be used. In the next section we will describe this procedure for fine tuning all examined verification methods of this study.

Experimental Study

Description of PAN Datasets

In the framework of PAN evaluation campaigns in authorship verification between 2013 and 2015 several corpora were built to cover multiple natural languages, genres and degrees of difficulty (Juola & Stamatatos, 2013; Stamatatos et al., 2014, 2015). These corpora are usually exploited to evaluate new authorship verification approaches. In this study, we used all datasets released in PAN-2014 and PAN-2015 shared tasks (the corresponding datasets from PAN-2013 are of very small size). A PAN dataset consists of a collection of verification problems, each problem comprises a small set of known documents, all by the same author, and exactly one unknown document. There are separate training and evaluation parts. All datasets are balanced regarding the number of positive (same-author) and negative (different-author) instances, in both training and evaluation parts. All documents within a problem are in the same language and text-length may vary. According to the specific dataset, genre and thematic area may be controlled (same for all documents within a problem) or can be mixed (cross-topic or cross-genre conditions) (Stamatatos et al., 2015).

Dataset	Language	Genre	Topic	#Problems	$\|D_{known}\|$	$\|d\|$
PAN14-DE	Dutch	essays	similar	96	2.0	398
PAN14-DR	Dutch	reviews	similar	100	1.0	116
PAN14-EE	English	essays	similar	200	2.6	833
PAN14-EN	English	novels	similar	200	1.0	6,104
PAN14-GR	Greek	articles	similar	100	2.7	1,537
PAN14-SP	Spanish	articles	similar	100	5.0	1,121
PAN15-DU	Dutch	mixed	mixed	165	1.7	360
PAN15-EN	English	plays	mixed	500	1.0	536
PAN15-GR	Greek	articles	mixed	100	2.8	756
PAN15-SP	Spanish	mixed	mixed	100	4.0	946

Table 1
Properties of PAN-2014 and PAN-2015 authorship verification evaluation datasets. $|D_{known}|$ *and* $|d|$ *are average number of known documents and average text-length (in words), respectively.*

Table 1 provides a description of properties for each evaluation dataset used in this study (the corresponding training datasets have similar properties). As can be seen, the dataset properties are not homogeneous. For example, the number of known documents may be minimum (e.g., PAN14-EN), of varying size (e.g., PAN14-GR), or fixed (e.g., PAN14-SP). A document may consist of just a couple of paragraphs (e.g., PAN14-DR), a couple of pages (e.g., PAN14-GR), or several pages (e.g., PAN14-EN). All PAN-2014 datasets control both genre and thematic area within a verification problem, while PAN-2015 datasets are more challenging including cross-genre (e.g., PAN15-DU) and cross-topic (e.g., PAN15-GR) cases where the topic/genre of documents in D_{known} is different from that of d_u. More details on these datasets are provided in (Stamatatos et al., 2014, 2015).

External Documents

Given that the extrinsic methods need a set of external documents for each verification problem, we follow the practice of previous studies to collect such a set for each evaluation corpus (Khonji & Iraqi, 2014; Seidman, 2013). In particular, we formed queries from the documents of the training corpus in each dataset, submitted these queries in a search engine, and downloaded impostor documents from the world wide web. In contrast to previous studies, our approach was only based on the known documents of the training corpus. Since the external documents are going to be used as competitors to the known documents, it makes sense to use information from the known documents to retrieve them. Moreover, in case we have a cross-topic or cross-genre dataset, then the properties of the known documents are quite different than the ones of the unknown documents.

In order to extract significant terms from the documents and form appropriate queries, we first performed clustering on the set of unique known documents D of the training corpus based on EM algorithm (Dempster, Laird, & Rubin, 1977). The resulting clusters C indicate basic thematic areas of those documents. Then, we extract the most significant terms from both documents and clusters. In more detail, the set of document-related terms $T_{document}$ contains all terms t occurring in a document $d \in D$ given that the following criteria are satisfied:

$$t \in T_{document} : \exists d \in D . f(t,d) \geq 2 \wedge df(t) \leq 3 \wedge |t| \geq 3 \tag{6}$$

where $f(t,d)$ is the raw frequency, $df(t)$ is the document frequency (the number of different documents the term occurs), and $|t|$ is the length (in characters) of term t. Practically, thematically-related terms are selected. In addition, the set of cluster-related terms $T_{cluster}$ comprises all terms t occurring in a cluster c when the following criteria apply:

$$t \in T_{cluster} : \exists c \in C . f(t,c) \geq 2 \wedge cf(t) = 1 \wedge |t| \geq 3 \tag{7}$$

where $cf(t)$ is the cluster frequency of t (the number of different clusters the term occurs). The extracted terms are characteristic of a certain cluster since they do not occur in other

clusters. Then, several queries composed by exactly five terms are formed by randomly drawing (without replacement) from $T_{document} \cup T_{cluster}$. We used *Bing* search engine and downloaded the first results of each query. These documents have mainly thematic similarities with the known documents. The downloaded documents were stripped from HTML tags and filtered out so that only relatively long documents (longer than the average text-length of the respective dataset) remain. The size of downloaded documents included in the set of external documents of each dataset ranges from 400 to 700. It should be noted that the impostor documents are truncated based on the length of known documents in each verification problem separately, so that their size does not affect the calculation of similarity score.

Experimental Setup

To find the most appropriate values for the required hyper-parameters of each of the examined methods, we performed grid search considering a range of possible values for each parameter and attempting to maximize performance on the PAN-2014 and PAN-2015 training dataset (Stamatatos et al., 2014, 2015). This process was repeated for each benchmark dataset separately. Since the participants of PAN-2014 and PAN-2015 were ranked according to the area under the receiver operating characteristic curve (AUC), we also use this evaluation measure.

First, as concerns the intrinsic approaches, the only hyper-parameter that should be tuned is the number of latent topics k. The examined range of values of k is $\{20, 30, 40, 50, 70, 100, 150, \ldots, 300\}$. Similarly, the number of topics is tuned for extrinsic methods as well. In addition, as can be seen in algorithms 1 and 2, there are several hyper-parameters than need to be set. Some previous studies fine-tune all these parameters separately (Khonji & Iraqi, 2014; Seidman, 2013). However, in this study we follow the approach described by Potha and Stamatatos (2017) to simplify this process. In particular, let a=$|Impostors_{problem}|$, then the number of repetitions and the number of impostors that are randomly selected in each repetition are defined as *repetitions*= a/5 and $|Impostors_{repetition}|$=a/10, respectively. Note that this analogy is not far away with the settings extracted when each hyper-parameter is tuned separately (Khonji & Iraqi, 2014;

Seidman, 2013). The following range of a values is examined for each dataset: $\{50, 100, \ldots, 300\}$. A fix $rate = 0.5$ is considered. Furthermore, in accordance to previous studies (Seidman, 2013), for EIP method, the aggregation function is selected among average, minimum, and maximum. Note also that the search engine used to collect the external documents (we use Bing) can also be viewed as a hyper-parameter.

Experimental Results

In all results, AUC of the receiver-operating characteristic curve is used as evaluation measure. This was also used in PAN-2014 and PAN-2015 shared tasks (Stamatatos et al., 2014, 2015). Taking into consideration that the extrinsic methods (EIB and EPB) are stochastic, each experiment is repeated five times and the average performance is reported. Moreover, the feature set contains all terms (words) occurring at least 5 times in the training corpus.

LSI vs. LDA. First, we compare the performance of the four examined models representing all basic author verification paradigms (i.e., IIB, IPB, EIB, EPB) when combined with either LSI or LDA topic modeling. Based on that analysis we can reveal if a certain topic modeling technique is more appropriate for specific verification paradigms or there is a common pattern in any combination of paradigm and topic modeling approach.

Table 2 presents the AUC scores of all 4 verification approaches when either LSI or LDA is used for all PAN-2014 and PAN-2015 evaluation datasets. Recall that parameter settings for each model have been optimized based on the corresponding training dataset. As can be seen, extrinsic methods are better than intrinsic ones and profile-based approaches are better than the corresponding instance-based ones. Moreover, LDA helps extrinsic methods to achieve improved results in comparison to LSI. On the other hand, the most successful intrinsic method (IPB) is better combined with LSI rather than LDA (Potha & Stamatatos, 2018). Clearly, the top performing method seems to be EPB that achieves best results in almost all datasets, especially when combined with LDA.

Given that parameter k is the only one directly associated with topic modeling techniques, we present a more analytical view of obtained performance of the examined

Dataset	IIB-LSI	IIB-LDA	IPB-LSI	IPB-LDA	EIB-LSI	EIB-LDA	EPB-LSI	EPB-LDA
PAN14-DE	0.708	0.907	0.960	0.924	0.963	0.953	0.976	**0.978**
PAN14-DR	0.674	0.616	0.686	0.560	0.724	0.740	0.750	**0.759**
PAN14-EE	0.578	0.450	0.567	0.425	0.695	0.703	0.726	**0.773**
PAN14-EN	0.684	0.660	0.696	0.753	0.771	0.768	0.780	**0.792**
PAN14-GR	0.728	0.754	**0.922**	0.867	0.905	0.901	0.918	0.921
PAN14-SP	0.732	0.746	0.890	0.758	0.833	0.837	0.852	**0.910**
PAN15-DU	0.657	0.454	0.639	0.372	0.726	0.730	0.734	**0.759**
PAN15-EN	0.699	0.766	0.811	0.788	0.753	0.801	0.848	**0.857**
PAN15-GR	0.516	0.616	0.836	0.717	0.858	0.894	0.877	**0.904**
PAN15-SP	0.732	0.800	0.750	0.828	0.909	0.936	0.945	**0.951**
Average	0.671	0.677	0.776	0.699	0.814	0.826	0.841	**0.860**

Table 2
AUC scores of the author verification methods on PAN datasets.

verification approaches when k varies. Figure 3 depicts the average performance of verification models over all 10 PAN-2014 and PAN-2015 datasets when k is fixed to a specific value in the range of 20 to 300 for all datasets. For extrinsic methods LDA is clearly better than LSI for the whole range of k values. Performance improves considerably by increasing the number of topics and all methods achieve best results when 200-250 topics are used. On the other hand, LSI is much better than LDA when IPB is concerned. Performance does not seem to be much affected by varying k in that case. However, the best results are achieved when a low number of topics is used. Finally, for IIB we get a mixed picture. For low number of topics, LDA is the best option. When k is more than 100, then LSI clearly outperforms LDA and achieves the best results.

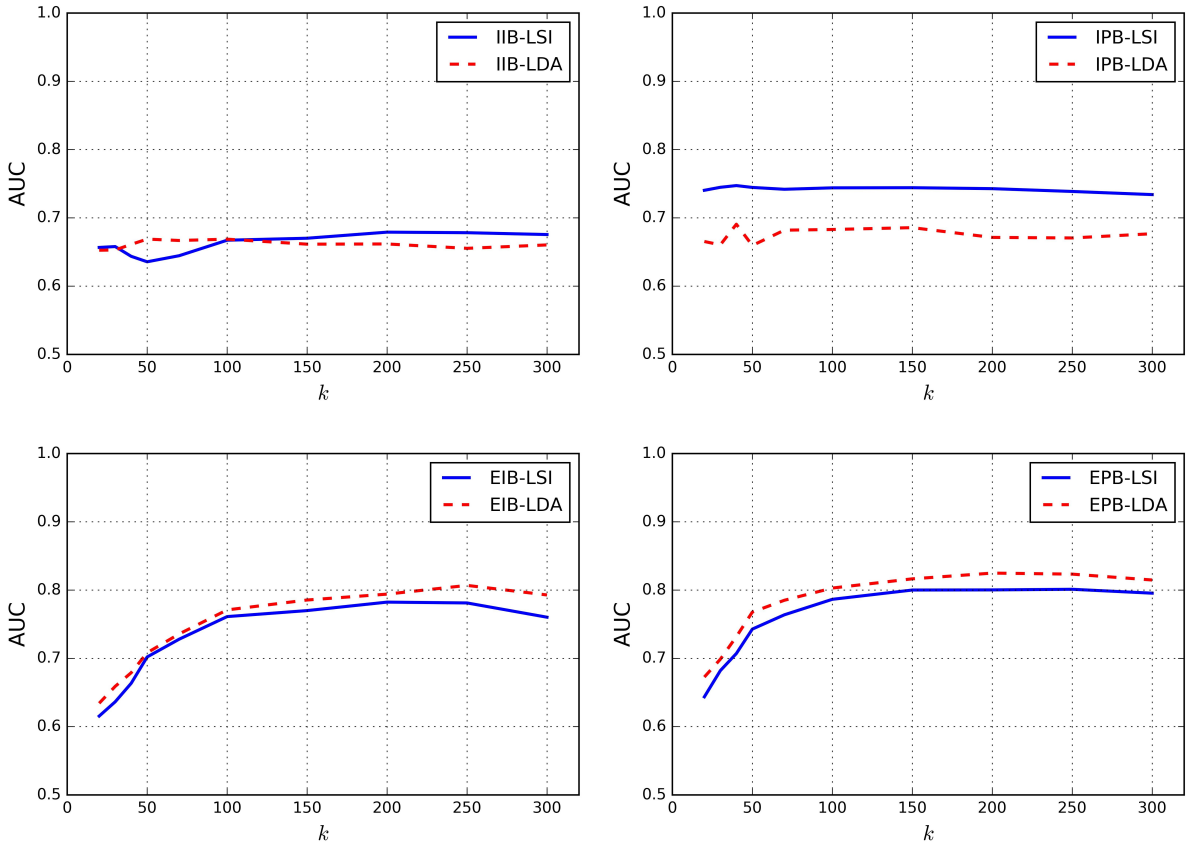

Figure 3. Average performance of IIB (upper left), IPB (upper right), EIB (lower left), and EPB (lower right) on PAN-2014 and PAN-2015 corpora when combined with either LSI or LDA and a varying number of latent topics.

Source of Topic Models. So far, the topic models have been extracted by using the set of documents in the training dataset. These are a few hundred documents for each dataset. In order to examine how the size of the document collection used to extract the topic models affects the performance of author verification approaches, we also conducted experiments using a larger collection. More specifically, we add the external documents collected for each dataset to the training documents and we use this enriched set of documents to extract the topic models for both LSI and LDA. In each training dataset the number of unique documents (several documents are used in more than one verification problem) ranges from 50 to 650 (roughly). Therefore, the enriched document collection is significantly larger in comparison to the training documents. The topic models extracted from the enriched collection of documents are referred as LSI* and LDA* to distinguish them from the ones obtained by the training texts only.

Table 3 demonstrates the improvement in the performance (difference of AUC scores) of the verification approaches when the topic models are extracted based on the enriched set of documents in comparison to the case where only the training texts are used (i.e., with respect to Table 2). Statistical significance of these differences is estimated using an approximate randomization test (Noreen, 1989). The null hypothesis is that there is no difference between the two cases and we reject this hypothesis when $p < 0.05$. As can be seen, in general, the models extracted from the enriched corpus are more effective. In addition, LDA models are more improved (in average) than the corresponding LSI ones (with the exception of IPB) and extrinsic models gain more than intrinsic ones. As concerns the individual datasets, the results on PAN14-EE are improved the most while the average results on PAN14-DE and PAN14-SP are slightly decreased. This does not seem to correlate with the relative increase in the number of documents used to extract the topic models.

Figure 4 depicts the performance of the four examined paradigms in combination with either LSI* or LDA* for a varying number of latent topics (k). This can be directly compared with Figure 3 and this comparison reveals the contribution of the enriched document collections to extract the latent topics. In general, the same patterns can be viewed in both figures. LDA* is better for the extrinsic models as well as for IIB while LSI* is clearly a better choice for IPB. A remarkable exception concerns IIB models where LSI* is better than LDA* for very low number of latent topics and it is outperformed by LDA* for larger sets of latent topics (essentially the opposite of the pattern depicted in Figure 3).

Comparison to the State of the Art. In all our experiments, we use the training set to tune the hyper-parameters of the verification methods and the performance on the evaluation datasets is measured by the area under the receiver-operating characteristic curve. This makes our reported results directly comparable to the ones obtained by PAN participants in PAN-2014 and PAN-2015 shared tasks (Stamatatos et al., 2014, 2015). Moreover, our results can directly be compared to the ones of other published methods that use exactly the same experimental settings and evaluation measures. The following state-of-the-art methods (ranked in chronological order) are used to estimate the competitiveness of the proposed methods:

- Jankowska et al. (2014): This is an intrinsic and instance-based verification method.

Dataset	IIB-LSI*	IIB-LDA*	IPB-LSI*	IPB-LDA*	EIB-LSI*	EIB-LDA*	EPB-LSI*	EPB-LDA*	*Average*
PAN14-DE	**0.006**	**0.004**	0.022	-0.104	0.009	**0.024**	-0.026	0.004	*-0.008*
PAN14-DR	**0.031**	**0.077**	**-0.040**	-0.109	0.029	0.033	-0.007	**0.029**	*0.005*
PAN14-EE	**-0.040**	**0.088**	**0.214**	**0.195**	**0.048**	**0.046**	**0.028**	0.014	*0.074*
PAN14-EN	**-0.105**	**0.137**	**0.065**	0.006	**0.002**	**0.035**	-0.008	**0.005**	*0.017*
PAN14-GR	0.002	0.020	**-0.003**	**-0.073**	**0.009**	**0.010**	**0.020**	**0.019**	*0.001*
PAN14-SP	-0.018	-0.038	**0.012**	**-0.130**	**0.032**	**0.047**	**0.057**	**0.001**	*-0.005*
PAN15-DU	**0.003**	**0.051**	**-0.050**	**0.221**	**-0.035**	**0.026**	0.002	**0.021**	*0.030*
PAN15-EN	**0.076**	**-0.083**	**-0.009**	**0.004**	**0.086**	**0.084**	0.033	**0.030**	*0.028*
PAN15-GR	0.118	**-0.017**	**0.015**	**0.053**	0.042	**-0.004**	**0.047**	**0.042**	*0.037*
PAN15-SP	0.051	**-0.067**	**0.170**	-0.018	0.012	**0.010**	0.001	**0.014**	*0.022*
Average	*0.012*	*0.017*	*0.040*	*0.005*	*0.023*	*0.031*	*0.015*	*0.018*	*0.020*

Table 3
Improvement in performance (difference in AUC) between methods using topic models extracted from the enriched collection and topic models extracted from training documents only. Statistically significant differences ($p < 0.05$) are indicated in boldface.

Due to its simplicity and lack of parameters to be tuned, it served as baseline in PAN-2014 and PAN-2015 shared tasks (Stamatatos et al., 2014, 2015). IIB was inspired by this method. However, it uses character n-gram features and a different similarity function.

- Fréry et al. (2014): This method applies a decision tree learning algorithm in author verification by viewing the training corpus as a collection of positive and negative instances of a binary classification task. It acquired the second top-ranked position in PAN-2014.

- Khonji and Iraqi (2014): This is a variant of the Impostors method that attained the

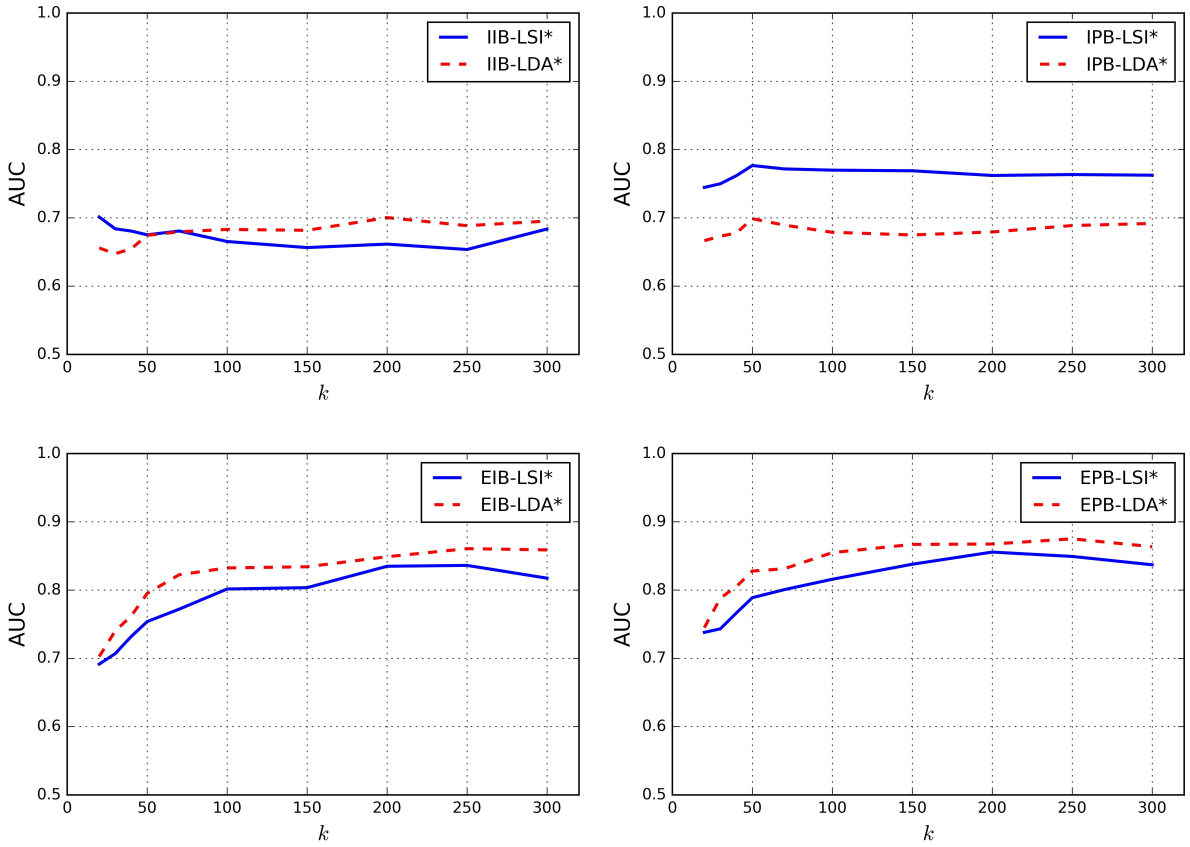

Figure 4. Average performance of IIB (upper left), IPB (upper right), EIB (lower left), and EPB (lower right) on PAN-2014 and PAN-2015 corpora when combined with either LSI* or LDA* and a varying number of latent topics.

overall top-ranked position in PAN-2014 shared task (Stamatatos et al., 2014). It is an extrinsic and instance-based method.

- META-PAN14 (Stamatatos et al., 2014): This a meta-model combining all 13 PAN-2014 submissions in an heterogeneous ensemble by averaging their answers. This meta-model was found more effective overall than any individual PAN participant method.

- Potha and Stamatatos (2014): This is an intrinsic and profile-based verification method based on character n-gram features. It is quite robust and achieved very good results on PAN-2013 datasets.

- Bagnall (2015): This is an extrinsic verification model based on a character-level multi-headed recurrent neural network language model. It applies both instance-based

and profile-based paradigms and it was the top-ranked submission in PAN-2015 (Stamatatos et al., 2015).

- Moreau et al. (2015). This is an ensemble approach that is based on several heterogeneous verification models. It achieved very good results, especially in the most challenging cross-genre PAN15-DU dataset, and was ranked second overall in PAN-2015 (Stamatatos et al., 2015).

- Pacheco et al. (2015): This an extrinsic and profile-based approach that combines a heterogeneous set of features, including LDA topics. It achieved the third overall position in PAN-2015 (Stamatatos et al., 2015).

- META-PAN15 (Stamatatos et al., 2015): This a meta-model combining all 18 PAN-2015 submissions in an heterogeneous ensemble. Although the main idea is similar to META-PAN14, this method was not equally effective since several PAN-2015 submissions produced very low results affecting the quality of the ensemble.

- Hernández and Calvo (2017): This is an intrinsic and instance-based method that utilizes topic modeling. LDA features are used to represent known and unknown documents and then the difference between them reveals a tendency to use roughly similar or different latent topics.

- Ding et al. (2018): This is an intrinsic and profile-based method that uses a joint neural network to learn distributed word representations as well as topical and lexical biases related to the global topic of the document and the personal bias of the author to use specific words given that topic. The reported results of this method are the best ones so far for the PAN-2014 datasets.

Note that the published results for some of the above methods only refer to either PAN-2014 or PAN-2015 datasets. Thus, we use a different set of baseline methods for these two sets of datasets. Tables 4 and 5 demonstrate the effectiveness of the state-of-the-art methods per dataset and in average for PAN-2014 and PAN-2015, respectively. In addition, we show the corresponding performance of IIB, EIB, and EPB combined with LDA* while IPB is

combined with LSI*. In all cases, the topic models are extracted from the enriched document collections. In addition, we also consider the combination of the two best-performing extrinsic methods (EIB-LDA* and EPB-LDA*) by averaging their answers for each verification problem. Note that since there are 5 runs for each of these models, we take the average of all 10 runs. This (EIB-LDA*+EPB-LDA*) is a very simple form of classifier ensemble inspired by META-PAN14 and META-PAN15 (Stamatatos et al., 2014, 2015).

Among the verification methods examined in this study, only IIB seems not to be highly competitive with state-of-the-art methods. However, it should be underlined that, in most of the cases, it outperforms the approach of Jankowska et al. (2014) from which it was inspired. This clearly demonstrates the contribution of topic modeling techniques to improve an existing author verification method. The other intrinsic method (IPB) is much more competitive. More specifically, it achieves very good results in datasets with increased number of known documents like PAN14-GR, PAN14-SP, PAN15-GR, and PAN15-SP (see table 1). Note that in these specific datasets EPB also outperforms EIB. This indicates that profile-based methods are better able to handle a relatively large size of D_{known}.

Clearly, the extrinsic methods examined in this study are the most effective ones, outperforming in some cases the best reported results so far. Given that the best state-of-the-art methods use much more sophisticated models based on representation learning (Ding et al., 2018) and neural network language models (Bagnall, 2015), it can be concluded that topic modeling provides an effective approach in author verification when it is fine-tuned and appropriately combined with suitable paradigms. The improved average results in the challenging PAN-2015 datasets indicate that the examined methods are not easily confused in cross-topic conditions and the extracted latent topics capture useful stylistic information. In addition, topic modeling techniques are fast and significantly reduce the dimensionality of document representation. This has a positive effect in the efficiency of the proposed methods.

The combination of EIB and EPB achieves the best average results in both PAN-2014 and PAN-2015 datasets. This verifies the conclusions of previous studies that ensembles of classifiers based on multiple, possibly heterogeneous models, can further improve the performance of individual verification methods (Moreau et al., 2015; Stamatatos et al., 2014).

	PAN14-DE	PAN14-DR	PAN14-EE	PAN14-EN	PAN14-GR	PAN14-SP	Average
Jankowska et al. (2014)	0.865	0.607	0.543	0.453	0.706	0.713	0.648
Fréry et al. (2014)	0.906	0.601	0.723	0.612	0.679	0.774	0.716
Khonji & Iraqi (2014)	0.913	0.736	0.590	0.750	0.889	0.898	0.797
META-PAN14 (2014)	0.957	0.737	0.781	0.732	0.836	0.898	0.824
Potha and Stamatatos (2014)	0.918	0.712	0.553	0.664	0.659	0.737	0.707
Hernández and Calvo (2017)	0.855	0.577	0.780	0.644	0.686	0.576	0.686
Ding et al. (2018)	**0.998**	0.658	**0.887**	0.767	0.924	**0.934**	0.876
IIB-LDA*	0.911	0.693	0.538	0.797	0.774	0.710	0.737
IPB-LSI*	0.982	0.646	0.781	0.761	0.919	0.902	0.832
EIB-LDA*	0.977	0.773	0.749	**0.803**	0.911	0.884	0.849
EPB-LDA*	0.982	0.788	0.787	0.797	0.940	0.911	0.867
EIB-LDA*+EPB-LDA*	0.980	**0.836**	0.765	0.799	**0.980**	0.921	**0.880**

Table 4
Comparison of state-of-the-art methods with the best methods of this study on PAN-2014 datasets.

The Effect of Genre of External Documents

So far, in all experiments the set of external documents required by the extrinsic methods are downloaded from the WWW. This means that the genre of these documents most probably do not match the one of the documents in question (d_u and D_{known}). In previous work, it is demonstrated that the performance of extrinsic methods can be considerably improved when the genre of all documents is the same (Koppel & Winter, 2014). In this section, we use another recently released corpus that will allow us to examine this effect. The *Enron authorship verification corpus* (EAV) (Halvani & Graner, 2018) includes emails by 80 authors extracted from the large pool of Enron email dataset (Klimt & Yang, 2004). For each

	PAN15-DU	PAN15-EN	PAN15-GR	PAN15-SP	Average
Jankowska et al. (2014)	0.506	0.654	0.641	0.656	0.614
Potha and Stamatatos (2014)	0.632	0.754	0.682	0.726	0.698
Bagnall (2015)	0.700	0.811	0.882	0.886	0.820
Moreau et al. (2015)	**0.825**	0.709	0.887	0.853	0.819
Pacheco et al. (2015)	0.763	0.822	0.773	0.908	0.817
META-PAN15 (2015)	0.696	0.786	0.779	0.894	0.754
Hernández and Calvo (2017)	0.751	0.853	0.709	0.783	0.774
IIB-LDA*	0.504	0.683	0.699	0.802	0.672
IPB-LSI*	0.589	0.802	0.851	0.920	0.791
EIB-LDA*	0.756	0.885	0.890	0.946	0.869
EPB LDA*	0.780	0.887	0.946	0.965	0.894
EIB-LDA*+EPB-LDA*	0.784	**0.902**	**0.966**	**0.965**	**0.904**

Table 5
Comparison of state-of-the-art methods with the best methods of this study on PAN-2015 datasets.

author there are at most 4 documents and each document includes an aggregation of email messages by that author (up to 5k characters). The email messages included in this corpus have been manually preprocessed to remove greetings, signatures, telephone numbers named entities etc. We follow exactly the same procedure as described by Halvani and Graner (2018) to split this corpus into training and test parts and form positive and negative author verification cases. The training part is balanced with 24 positive and 24 negative verification cases while the test part includes 56 positive cases and 3,080 negative cases.

The parameters of the methods presented in this study were estimated based on the training part of the corpus as described in Experimental Setup section. As concerns the set of

	Genre-agnostic				Email messages			
Method	**LSI**	**LSI***	**LDA**	**LDA***	**LSI**	**LSI***	**LDA**	**LDA***
Halvani et al. (2017)	0.909							
Seidman (2013)	0.896							
EIB	0.892	0.913	0.907	0.932	0.885	0.930	0.944	0.967
EPB	0.919	0.920	0.936	0.957	0.924	0.924	0.964	**0.973**
EIB+EPB	0.932	0.942	0.956	0.944	0.913	0.949	0.968	0.972

Table 6
AUC results of extrinsic methods on the EAV corpus using either genre-agnostic external documents or email messages and baselines performance as reported by Halvani and Graner (2018).

external documents required by the extrinsic methods, we explored two alternatives. First, we follow the approach used in the previous experiments, namely external documents are downloaded from the WWW by using the Bing search engine. In total, 490 documents are collected and we call this *genre-agnostic* set of external documents. Second, we use the PAN-2011 authorship identification corpus (Argamon & Juola, 2011), also based on email messages extracted from the Enron dataset, to form impostor documents. We concatenate email messages by the same author so that each impostor document to have similar length with the documents of the EAV corpus. However, we do not apply any other preprocessing of email messages. A set of 410 external documents (email messages) is obtained. In addition, similar to the previous experiments, we examine two cases of extracting the latent topics: one using only the training texts (based on either LSI or LDA) and another using the enriched set of training and external texts (LSI* and LDA*). In the case of genre-agnostic external documents, the enriched collection comprises a mix of genres.

Table 6 shows the AUC evaluation results of the extrinsic verification methods examined in this study on the EAV corpus. As baselines, we use the best results reported by Halvani and Graner (2018) on the same corpus. In more detail, the best results so far in this corpus are obtained by a profile-based and intrinsic verification method based on text

compression (Halvani et al., 2017) and the GI method (Seidman, 2013). As expected, the set of email impostors assists extrinsic methods to achieve higher results in comparison to the case genre-agnostic external documents are used. This difference is more evident when LDA (and LDA*) is applied. However, the extrinsic verification methods based on the genre-agnostic set of external documents are competitive and they all clearly surpass the baselines when they are based on enriched LDA* models. In almost all cases, the use of enriched models (LSI* and LDA*) seems to enhance the performance of verifiers. When genre-agnostic external documents are concerned, this shows that information in documents belonging to a mix of genres can be useful to define latent topics.

Conclusions

This paper provides the first systematic study of the contribution of topic modeling techniques to improve author verification methods. We combined well-known topic modeling techniques with author verification methods that cover main paradigms in this field, namely, intrinsic and instance-based, intrinsic and profile-based, extrinsic and instance-based, and extrinsic and profile-based methods. It was demonstrated that in all cases, topic modeling can improve the effectiveness of the methods.

Profile-based methods seem to be better able to take full advantage of topic modeling. This is especially evident when multiple documents of known authorship are available. In addition, extrinsic methods are more effective than intrinsic methods, as it has been already indicated by previous studies (Juola & Stamatatos, 2013; Stamatatos et al., 2014, 2015). We demonstrated here that this is also true when topic modeling techniques are used. A new extrinsic and profile-based author verification algorithm that exploits topic modeling (EPB) has been used and its performance is highly competitive when compared with state-of-the-art methods.

The most successful author verification methods so far are based on sophisticated approaches using representation learning (Ding et al., 2018) and neural network language models (Bagnall, 2015). The methods presented in this study are much simpler, efficient, and equally effective. The presented results show that topic modeling techniques when

appropriately combined with suitable paradigms can provide very effective author verification systems.

LDA seems to be more effective than LSI for the extrinsic methods. Moreover, LDA models are more improved than LSI models when an enriched corpus is used to extract the latent topics. It remains to be seen if a greater improvement can be expected when even larger and more heterogeneous document collections are used to extract topic models. On the other hand, LSI is clearly more suitable for IPB, that seems to be the most effective intrinsic method examined in this study.

In order to achieve enhanced performance, the genre of external documents should match the genre of the documents in a verification case. However, if this is not cost-effective or even not possible, we demonstrated that easily collected genre-agnostic external documents can provide competitive performance.

The combination of the two most effective extrinsic verification methods, using LDA and latent topics extracted from an enriched document collection, outperforms all individual methods and the best state-of-the-art methods in average for both PAN-2014 and PAN-2015 datasets. This demonstrates the potential of ensembles of author verification methods. The development of more sophisticated ensembles exploiting topic modeling techniques to achieve high diversity is an open research direction.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163–222). Springer Berlin Heidelberg.

Argamon, S., & Juola, P. (2011). Overview of the international authorship identification competition at PAN-2011. In V. Petras, P. Forner, & P. Clough (Eds.), *Notebook Papers of CLEF 2011 Labs and Workshops.*

Bagnall, D. (2015). Author Identification using multi-headed Recurrent Neural Networks. In L. Cappellato, N. Ferro, J. Gareth, & E. San Juan (Eds.), *Working Notes Papers of the CLEF 2015 Evaluation Labs.*

Barbon, S., Igawa, R., & Bogaz Zarpelão, B. (2017). Authorship verification applied to detection of compromised accounts on online social networks: A continuous approach. *Multimedia Tools and Applications*, *76*(3), 3213-3233.

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

Brocardo, M., Traore, I., Woungang, I., & Obaidat, M. (2017). Authorship verification using deep belief network systems. *International Journal of Communication Systems*, *30*(12).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

Ding, S., Fung, B., Iqbal, F., & Cheung, W. (2018). Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*.

Duman, S., Kalkan-Cakmakci, K., Egele, M., Robertson, W., & Kirda, E. (2016). Emailprofiler: Spearphishing filtering with header and stylometric features of emails. In *Proceedings - international computer software and applications conference* (Vol. 1, p. 408-416).

Fréry, J., Largeron, C., & Juganaru-Mathieu, M. (2014). Ujm at clef in author identification.

Proceedings CLEF-2014, Working Notes, 1042–1048.

Halvani, O., & Graner, L. (2018). Rethinking the evaluation methodology of authorship verification methods. In P. Bellot et al. (Eds.), *Proceedings of the international conference of the cross-language evaluation forum for european languages* (pp. 40–51). Springer International Publishing.

Halvani, O., Graner, L., & Vogel, I. (2018). Authorship verification in the absence of explicit features and thresholds. In *European conference on information retrieval* (pp. 454–465).

Halvani, O., Winter, C., & Graner, L. (2017). On the usefulness of compression models for authorship verification. In *Proceedings of the 12th international conference on availability, reliability and security* (pp. 54:1–54:10). New York, NY, USA: ACM. doi: 10.1145/3098954.3104050

Halvani, O., Winter, C., & Pflug, A. (2016). Authorship verification for different languages, genres and topics. *Digital Investigation*, *16*, S33-S43.

Hernández, C. Á., & Calvo, H. (2017). Author verification using a semantic space model. *Computación y Sistemas*, *21*(2).

Hürlimann, M., Weck, B., van den Berg, E., Šuster, S., & Nissim, M. (2015). GLAD: Groningen Lightweight Authorship Detection. In L. Cappellato, N. Ferro, G. Jones, & E. San Juan (Eds.), *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers.* CEUR-WS.org.

Jankowska, M., Milios, E., & Keselj, V. (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* (pp. 387–397).

Juola, P., & Stamatatos, E. (2013). Overview of the author identification task at PAN 2013. In *Working notes for CLEF 2013 conference.*

Kestemont, M., Luyckx, K., Daelemans, W., & Crombez, T. (2012). Cross-genre authorship verification using unmasking. *English Studies*, *93*(3), 340–356.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, *63*, 86-96.

Khonji, M., & Iraqi, Y. (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.

Klimt, B., & Yang, Y. (2004). The enron corpus: A new dataset for email classification research. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Machine learning: Ecml 2004* (pp. 217–226). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kocher, M., & Savoy, J. (2017). A simple and efficient algorithm for authorship verification. *Journal of the Association for Information Science and Technology*, *68*(1), 259-269.

Koppel, M., Schler, J., Argamon, S., & Winter, Y. (2012). The fundamental problem of authorship attribution. *English Studies*, *93*(3), 284-291.

Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, *8*, 1261–1276.

Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the American Society for Information Science and Technology*, *65*(1), 178-187.

Layton, R., Watters, P., & Ureche, O. (2013). Identifying faked hotel reviews using authorship analysis. In *Proceedings - 4th cybercrime and trustworthy computing workshop, ctc 2013* (p. 1-6).

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., & Stamatatos, E. (2015). Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, *89*, 134 - 147.

Moreau, E., Jayapal, A., Lynch, G., & Vogel, C. (2015). Author verification: Basic stacked generalization applied to predictions from a set of heterogeneous learners-notebook for pan at clef 2015. In L. Cappellato, N. Ferro, J. Gareth, & E. San Juan (Eds.), *Working Notes Papers of the CLEF 2015 Evaluation Labs.*

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys*, *50*(6).

Noreen, E. (1989). *Computer-intensive methods for testing hypotheses: An introduction*. Wiley.

Pacheco, M. L., Fernandes, K., & Porco, A. (2015). Random forest with increased generalization: A universal background approach for authorship verification. In L. Cappellato, N. Ferro, J. Gareth, & E. San Juan (Eds.), *Working Notes Papers of the CLEF 2015 Evaluation Labs.*

Panicheva, P., Cardiff, J., & Rosso, P. (2010). Personal sense and idiolect: Combining authorship attribution and opinion analysis. In *Proceedings of the international conference on language resources and evaluation, LREC.*

Potha, N., & Stamatatos, E. (2014). A profile-based method for authorship verification. In *Artificial intelligence: Methods and applications - proceedings of the 8th Hellenic conference on AI, SETN* (pp. 313–326).

Potha, N., & Stamatatos, E. (2017). An improved impostors method for authorship verification. *Proc. of the 8th International Conference of the CLEF Association, CLEF 2017, Lecture Notes in Computer Science*, *10456*, 138-144.

Potha, N., & Stamatatos, E. (2018). Intrinsic author verification using topic modeling. In *Artificial intelligence: Methods and applications - proceedings of the 10th hellenic conference on ai, SETN.*

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., ... Stamatatos, E. (2017). Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, *12*(1), 5–33.

Sanderson, C., & Guenter, S. (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* (pp. 482–491).

Sari, Y., & Stevenson, M. (2015). A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification. In L. Cappellato, N. Ferro, G. Jones, & E. San Juan (Eds.), *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers.* CEUR-WS.org.

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

Seidman, S. (2013). Authorship Verification Using the Impostors Method. In P. Forner, R. Navigli, & D. Tufis (Eds.), *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers.*

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship attribution with topic models. *Computational Linguistics*, *40*(2), 269–310.

Shahid, U., Farooqi, S., Ahmad, R., Shafiq, Z., Srinivasan, P., & Zaffar, F. (2017). Accurate detection of automatically spun content via stylometric analysis. In *2017 ieee international conference on data mining (icdm)* (p. 425-434). doi: 10.1109/ICDM.2017.52

Shams, R., & Mercer, R. E. (2016). Supervised classification of spam emails with natural language stylometry. *Neural Computing and Applications*, *27*(8), 2315–2331. doi: 10.1007/s00521-015-2069-7

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, *60*, 538–556.

Stamatatos, E. (2016). Authorship verification: A review of recent advances. *Research in Computing Science*, *123*, 9–25.

Stamatatos, E. (2018). Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, *69*(3), 461-473.

Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B. (2015). Overview of the author identification task at PAN 2015. In L. Cappellato, N. Ferro, J. Gareth, & E. San Juan (Eds.), *Working Notes Papers of the CLEF 2015 Evaluation Labs.*

Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., ... Barrón-Cedeño, A. (2014). Overview of the author identification task at PAN 2014. In *CLEF working notes* (pp. 877–897).

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*(4), 471–495.

Stein, B., Koppel, M., & Stamatatos, E. (Eds.). (2007). *SIGIR 2007 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN).*

CEUR-WS.org.

Stein, B., Lipka, N., & z. Eissen, S. M. (2008). Meta analysis within authorship verification. In *Proceedings of the 19th international workshop on database and expert systems applications* (p. 34-39).

Stover, J. A., Winter, Y., Koppel, M., & Kestemont, M. (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), 239–242.

Tuccinardi, E. (2017). An application of a profile-based method for authorship verification: Investigating the authenticity of Pliny the Younger's letter to Trajan concerning the Christians. *Digital Scholarship in the Humanities*, *32*(2), 435-447.

Vaz, P. C., Martins de Matos, D., & Martins, B. (2012). Stylometric relevance-feedback towards a hybrid book recommendation algorithm. In *Proceedings of the fifth acm workshop on research advances in large digital book repositories and complementary media* (pp. 13–16). New York, NY, USA: ACM.