A Profile-Based Method for Authorship Verification

Nektaria Potha and Efstathios Stamatatos

Dept. of Information and Communication Systems Eng.
University of the Aegean
83200 – Karlovassi, Greece
pothanektaria@hotmail.com, stamatatos@aegean.gr

Abstract. Authorship verification is one of the most challenging tasks in style-based text categorization. Given a set of documents, all by the same author, and another document of unknown authorship the question is whether or not the latter is also by that author. Recently, in the framework of the PAN-2013 evaluation lab, a competition in authorship verification was organized and the vast majority of submitted approaches, including the best performing models, followed the instance-based paradigm where each text sample by one author is treated separately. In this paper, we show that the profile-based paradigm (where all samples by one author are treated cumulatively) can be very effective surpassing the performance of PAN-2013 winners without using any information from external sources. The proposed approach is fully-trainable and we demonstrate an appropriate tuning of parameter settings for PAN-2013 corpora achieving accurate answers especially when the cost of false negatives is high.

1 Introduction

Nowadays, text categorization provides effective solutions for handling the huge volumes of electronic text produced in Internet media [1]. 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). Authorship analysis attracts constantly increasing attention due to the large potential of important applications in intelligence (e.g., linking terrorist proclamations), security (e.g., verifying the identity of a person using a system), civil law (e.g., solving copyright disputes) etc.

Authorship attribution is the identification of the true author of a document given samples of undisputed documents from a set of candidate authors and has a long research history [6, 8, 20]. There are three main forms of this task usually examined in the relevant literature:

- *Closed-set attribution*: The set of candidate authors surely includes the true author of the questioned documents. This is the easiest version of the problem and most studies have focused on this, providing encouraging results. It should be noted that it is not an unrealistic scenario since in many forensic applications the investigators are able to filter out most of the persons involved in a case and produce a closed-set of suspects.

A. Likas, K. Blekas, and D. Kalles (Eds.): SETN 2014, LNCS 8445, pp. 313–326, 2014.

- *Open-set attribution*: The set of candidate authors may not contain the true author of some of the questioned documents. This is a much more difficult task especially when the size of the candidate set is small [12]. This setting fits all kind of applications including cases where anyone can be the true author of a questioned document (e.g., identifying the person behind a post in a blog).
- *Authorship verification*: This may be seen as a special case of open-set attribution where the set of candidate authors is singleton. As mentioned earlier, small candidate sets in open-set attribution are hard to be solved. All authorship attribution cases can be transformed to a set of separate authorship verification problems. So, the ability of a method to deal effectively with this fundamental task is crucial.

Very recently, there have been attempts to focus on fundamental problems of authorship attribution. Koppel et al. discuss the problem of determining if two documents are by the same author [13, 14]. This is a special case of the authorship verification task where the set of documents by the candidate author is singleton. In the PAN-2013 evaluation lab [9], a competition in authorship verification was organized where each verification problem consisted of a set of (up to 10) documents of known authorship by the same author and exactly one questioned document. The study of various attribution methods in such fundamental problems enables us to extract more general conclusions about their abilities and properties.

All authorship attribution methods fall under one of the following basic paradigms:

- *Instance-based paradigm*: All available samples by one author are treated separately. Each text sample has its own representation. Since these approaches are usually combined with discriminative machine learning algorithms, like support vector machines, they require multiple instances per class. Hence, when only one document is available for a candidate author, this document has to be split into multiple samples.
- *Profile-based paradigm*: All available text samples by one candidate author are treated cumulatively, that is they are concatenated in one big document and then a single representation is extracted to become the profile of the author.

In general, the former is more effective when multiple documents per author are available or when long documents (that can be split into multiple samples) are available. On the other hand, the profile-based paradigm is more effective when only short and limited samples of documents are available. Despite these advantages that are crucial when only one or two documents of known authorship are available, in PAN-2013 evaluation campaign 17 out of 18 participants followed the instance-based paradigm [9]. The only profile-based submission was ranked at the 11th position [2]. Therefore, it seems that instance-based approaches are more appropriate for authorship verification.

In this paper we claim the opposite. We present an authorship verification method following the profile-based paradigm and apply this method to the corpora produced in the framework of PAN-2013 using exactly the same evaluation setting. We provide evidence that profile-based authorship verification can be very effective surpassing the best performing submissions of that competition. The proposed approach is fully-trainable.

We show how the parameters of our method can be tuned given a training corpus so that the proposed method to be effectively applied to different natural languages and genres.

The rest of this paper is organized as follows: Section 2 presents previous work in authorship verification while Section 3 describes the proposed profile-based method. In Section 4 the experiments performed using the PAN-2013 corpora 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. [18]. They proposed an attribution model based on stylometric features extracted from an NLP tool and used multiple regression to produce the response function for a given author. Then, a threshold value (defined as a function of the multiple correlation coefficient) determines whether or not a questioned document was written by the examined author. This model was applied to a corpus of newspaper articles in (Modern) Greek providing good false acceptance rates and moderate false rejection rates.

A seminal authorship verification approach was proposed in [11]. The so-called *unmasking* method builds an SVM classifier to distinguish an unknown text from the set of known documents (all by a single author). Then, it removes a predefined amount of the most important features and iterates this procedure. If the drop in classification accuracy is not high, then the unknown document was written by the examined author. The logic behind this method is that at the beginning it will always be possible for the classifier to distinguish between the texts. When the the texts are by the same author, the differences will be focused on very specific features while when the texts are not by the same author the differences will be manifold. After the removal of some important features, texts by the same author will be difficult to be distinguished while in the opposite case, it will continue to be relatively easy to find other differences among them. The unmasking method is very effective when long documents are available since the unknown document has to be segmented into multiple pieces to train the SVM classifier. Its application to books was exceptional [11]. However, if only short documents are available, this method fails [16].

More recently, Koppel and Winter proposed the *impostors* method to determine whether two documents were by the same author [14]. This method first finds documents of similar genre and topic in the Web (the so-called impostors) and then it builds an ensemble model to verify whether one of the given documents is more similar to the other given document (same author) or one of the impostors (different author). Essentially, this method attempts to transform authorship verification from a one-class classification problem (i.e., the class of documents by a certain author) to a multi-class classification problem by introducing additional classes using documents found in external sources (i.e., the Web) and achieves very good results. However, since this process is automated, there is always the danger of retrieving a document that accidentally is by the same author with the documents of questioned authorship.

In the PAN-2013 evaluation lab, an authorship verification competition was organized [9]. The produced corpora covered three natural languages (i.e., English, Greek, and Spanish) and consist of a set of verification problems. Each problem provides a set of up to 10 documents by a single author and exactly one questioned document. In total, 18 teams participated in this competition. In general, the participant verification models can be distinguished into two main categories [9]:

- *Intrinsic verification models*: They are based exclusively on the set of documents of known documents by the same author and the questioned documents. They face the verification task as a one-class classification problem. Typical approaches of this category are described in [5, 7, 15].
- *Extrinsic verification models*: In addition to the given set of documents of known and unknown authorship, they use additional documents from external sources. They face the verification task as a multi-class classification problem. The winner participant, a modification of the impostors method, followed this approach [17]. Other similar approaches are described in [22-23].

The organizers of the evaluation campaign also reported the performance of a simple meta-model combining all the submitted outputs [9]. That heterogeneous ensemble had the best overall performance in both binary answers and real scores.

As concerns text representation, all kinds of features already studied in authorship attribution can also be used in authorship verification. At PAN-2013, the participants mainly used character 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.) that are also language-independent. The use of more sophisticated syntactic and semantic features doesn't seem to offer a significant advantage in this task possibly due to the low accuracy of the tools used to extract such features [9].

An important aspect is the appropriate parameter tuning of a verification model. Especially when the corpus comprises texts coming from different genres and natural languages, the verification model could be fine-tuned for each language/genre separately to improve its performance [7, 17]. According to each particular verification method, the parameters can be the type of used features, the number of used features, the threshold value used to produce the final decision, etc.

3 The Proposed Method

The method examined in this paper is a modification of the *Common N-Grams* (CNG) approach originally proposed by Keselj et al. [10] for closed-set attribution and later modified by Stamatatos [19]. Following, the profile-based paradigm, this method first concatenates all samples of known authorship into a single document and then extracts a character *n*-gram representation vector from this big document to serve as the author profile. Another vector is produced from the questioned document and the two vectors are compared using a dissimilarity function. If the resulting score is above

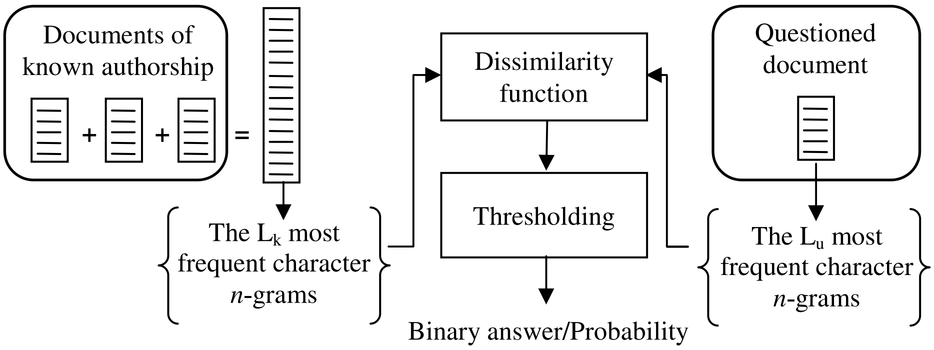

Fig. 1. The proposed profile-based approach for authorship verification

a certain threshold the questioned document is assigned to the author of known documents. This process is depicted in Figure 1. The original CNG approach uses profiles of the same length [10]. The modification of Stamatatos [19] uses assymetric profiles where the profile of the unknown text has the maximum possible length while the profile of the known texts by one candidate author is pre-defined. In our approach, the profile lengths of known and unknown documents are parameters to be set.

In total, the proposed approach has 4 parameters to be tuned. The first is the order of character n-grams (n). The second is the profile size of the questioned document (L_u), while the third refers to the corresponding profile size of the documents of known authorship (L_k). The last one is the dissimilarity function (d) discussed in Section 3.1.

Three PAN-2013 participants were also based on modifications of CNG. Jankowska, et al. [7] and Layton, et al. [15] modified this method to follow the instance-based paradigm, that is they produce separate representation vectors for each document of known authorship. The method of Jankowska, et al. requires at least two documents of known authorship [7]. Therefore in problems with just one document of known authorship they split it into two segments. The method described in [2] is more similar to ours since it also follows the profile-based paradigm. However, in our method we extract appropriately-tuned profile lengths for the questioned and known documents. Moreover, we examine a wider range of parameter values and select appropriate language-specific parameter settings.

3.1 Dissimilarity Function

Given two documents x and y and their profiles $P(x)$ and $P(y)$ (i.e., the sets of the most frequent character n-grams), the original CNG method [10] used a symmetrical dissimilarity function described as follows:

$$d_0(P(x),P(y)) = \sum_{g \in P(x) \cup P(y)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2 \quad (1)$$

where $f_x(g)$ and $f_y(g)$ are the normalized frequencies of occurrence of the character n-gram g in documents x and y, respectively. This function is very effective when the profile sizes are of similar size. However, when one profile is much shorter than the other, this measure becomes unstable and unreliable for closed-set attribution. An alternative and stable function in imbalanced conditions was introduced in [19]:

$$d_1(P(x), P(y)) = \sum_{g \in P(x)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2 \tag{2}$$

This is not a symmetrical function since it assumes that the first document is the questioned one and possibly shorter or much shorter than the second document of known authorship. Another alternative measure used in the Source Code Author Profiling (SCAP) method with very good results is the simplified profile intersection (SPI):

$$SPI(P(x), P(y)) = |P(x) \cap P(y)| \tag{3}$$

That is the mere counting of common character n-grams in both profiles. Note that *SPI* is a similarity function while d_0 and d_1 are dissimilarity functions. To have comparable dissimilarity measures one can use 1 – *SPI*(P(x),P(y)). In this study we used normalized versions of d_0, d_1, and *SPI* measures, as the one described in [21].

3.2 Production of Binary Answers and Probability Estimates

Having a dissimilarity score is not enough in authorship verification. We need a binary answer: a positive one in case the questioned document is estimated to be by the same author or a negative one in case it is estimated the opposite. In addition, we need a probability score for a positive answer to show the degree of certainty of that estimation. To produce binary answers, the most common approach is the definition of a threshold value. Any problem with score more than that threshold is considered to be a positive case. Usually, the definition of such threshold values depends on the training corpus [7, 17].

In this study, we use a simple thresholding procedure. Based on the dissimilarity scores produced for the problems of the training corpus that belong to the same genre/language we scale these values to the set [0,1] inclusive. Then, we use the same scaling function for every given evaluation problem belonging to the same genre/language. That way, the resulting score can be seen as a probability estimation of a negative answer (since we originally have dissimilarity rather than similarity scores). Its complementary value corresponds to the probability estimate of a positive answer. Let x be a verification problem, *score*(x,*dissimFunction*) be the dissimilarity score for this problem based on *dissimFunction,* and Y be a set of training verification problems of similar genre/language. Then, the probability estimate of a positive answer is expressed as:

$$p_+(x) = 1 - scale(score(x, dissimFunction), Y) \tag{4}$$

Table 1. Statistics of the PAN-2013 authorship verification corpus

	#problems	#documents	#characters (thousands)
Training	**35**	**189**	**1,535**
- English	10	42	265
- Greek	20	130	1,204
- Spanish	5	17	65
Test	**85**	**435**	**3,211**
- English	30	157	977
- Greek	30	178	1,714
- Spanish	25	100	520

Then, according to the percentage of positive/negative problems in the training corpus, we estimate the threshold value. For example, in a balanced corpus with 50% positive and 50% negative verification problems (as the one used in PAN-2013 competition), a "positive" binary answer (same author) is assigned to any verification problem x with $p_+(x)$>0.5. All problems with probability score lower than 0.5 will get the binary value "negative" (different author). Finally, all problems with score equal to 0.5 will remain unanswered, given that this option is allowed (as happened in the PAN-2013 competition).

4 Experimental Study

4.1 The PAN-2013 Evaluation Setting

In the framework of PAN-2013, an authorship verification corpus was built and released in early 2013 [9]. It includes a set of separate verification problems, each problem provides a set of up to 10 documents of known authorship, all by the same author, and exactly one questioned document. The corpus is segmented into a training part and an evaluation part. The latter was used for the final ranking of the participants and was released after the end of the submission deadline.

Three natural languages are represented in the corpus: English, Greek, and Spanish. The English part includes extracts from published textbooks on computer science and related disciplines. The Greek part contains opinion articles from a weekly newspaper while the Spanish part includes excerpts from newspaper editorials and short fiction. The PAN-2013 organizers report that the Greek part of the corpus is more challenging since they used stylometric techniques to match documents by different authors and find stylistically different documents by the same author. The language of each problem is encoded in its code name.

Table 1 shows some statistics of this corpus. As can be seen, the Greek part has more and longer documents while the Spanish part is under-represented especially in

Table 2. Global and local parameter settings of our approach extracted from the PAN-2013 training corpus

	L_u	L_k	n	d
Global	10,000	2,000	5	d_1
Local				
- English	1,000	1,000	5	d_1
- Greek	10,000	2,000	5	d_1
- Spanish	10,000	2,000	5	d_1

the training corpus. The latter makes the estimation of appropriate parameter settings for the Spanish part very difficult.

The PAN-2013 participants were asked to produce a binary YES/NO answer for each problem (corresponding to same author or different author) and, optionally, a probability estimate of a positive answer. Submissions were ranked based on recall and precision of correct answers combined by the (micro-average) F_1 measure. In addition, the participants that also produced probability estimates were ranked according to the area under the receiver operating characteristic curve (ROC-AUC) [9]. In this paper, we follow exactly the same evaluation settings to achieve compatibility of comparison with previously reported results.

4.2 Experiments

To find the most appropriate values for the 4 parameters of our method we examined a range of possible values and extracted the best models based on their performance on the PAN-2013 training set. We used ROC-AUC as the evaluation criterion. The following range of values were examined: $L_k \in \{1{,}000, 2{,}000, \ldots, 20{,}000\}$, $L_u \in \{1{,}000, 2{,}000, 10{,}000\}$, $n \in \{3,4,5\}$, and $d \in \{d_0, d_1, SPI\}$ as defined in formulas (1), (2), and (3). We first examined the entire training set and extracted global parameter settings. Then, the language information was considered and local parameter settings were produced for each one of the three languages.

Global Settings. The extracted global parameter settings, where the whole training corpus was considered, can be seen in Table 2. Figure 2 shows the AUC of authorship verification models based on different dissimilarity functions and the range of values of L_k when L_u=10,000 and n=5 for the full PAN-2013 training and test corpora. In both training and test corpora the basic patterns are the same. The best and more stable dissimilarity function is d_1. In addition, d_0 is competitive only for small values of L_k while its performance is negatively affected by increasing L_k. On the other hand, *SPI* achieves its best performance in around L_k=8,000. After that point its performance is similar with that of d_1. The best performing model for the training corpus (d=d_1, L_k=2,000) may not be the best performing model for the test corpus but it is very close to that.

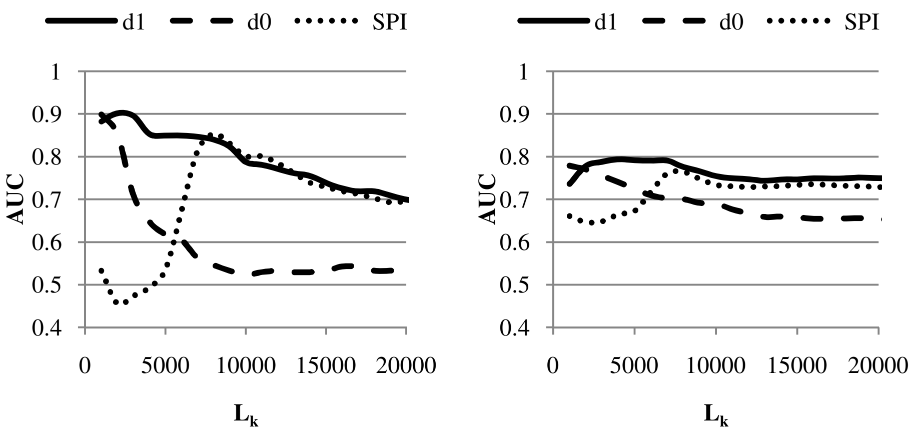

Fig. 2. The performance (AUC) of the proposed verification models on the training (left) and test (right) corpora with L_u=10,000 and n=5 and different dissimilarity functions

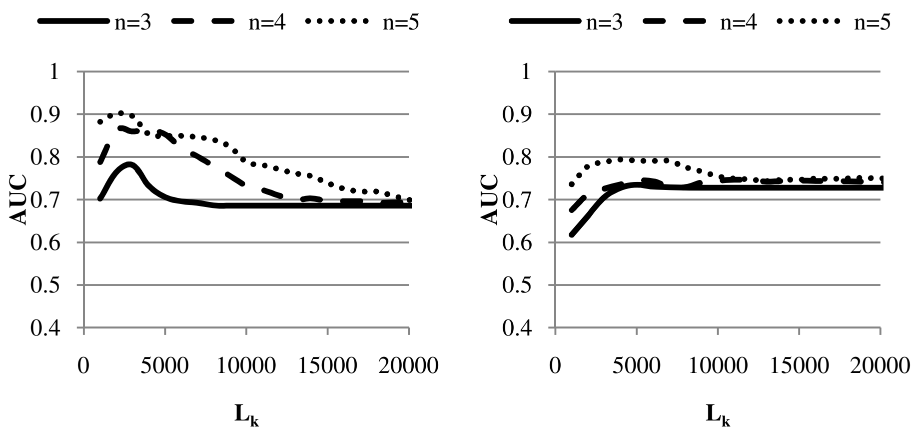

Fig. 3. The performance (AUC) of the proposed verification models on the training (left) and test (right) corpora with L_u=10,000 and d=d_1 and different orders of character n-grams

Figure 3 depicts the AUC scores of the verification models on the training and test sets based on L_u=10,000, d=d_1 and different values of n and L_k. Long character n-grams (n=5) seem to be the best and more stable option. On the other hand short n-grams (n=3) perform poorly. The same pattern applies to both training and test corpora.

Figure 4 shows the AUC scores of the verification models on the training and test sets based on d=d_1 and n=5 for different values of L_k and L_u. Apparently, increasing L_u helps to improve performance. For L_u>7,000 the performance is stabilized. This means that from the document of unknown authorship, all possible character n-grams

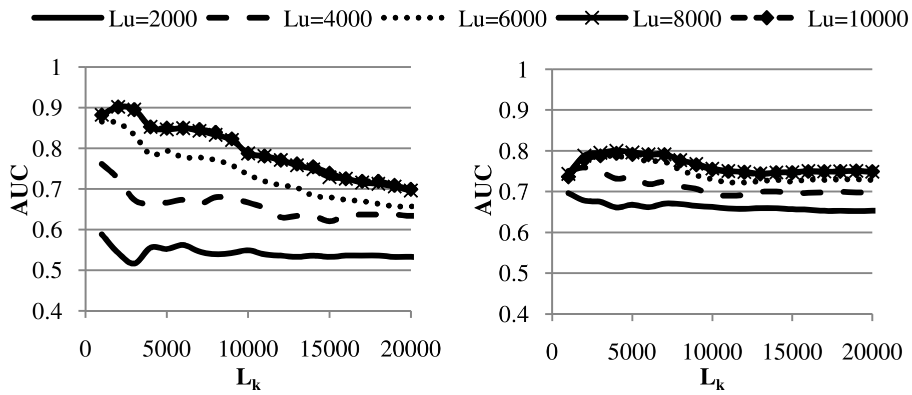

Fig. 4. The performance (AUC) of the proposed verification models on the training (left) and test (right) corpora with d=d_1, n=5, and different sizes of L_u

should be included in the verification model. This is not the case with the documents of known authorship. It seems that relatively low values of L_k (lower than 5,000) help achieving the best performance. In other words, from the documents of known authorship only the most frequent character n-grams should be included in the verification model. These patterns are consistent in both training and test sets.

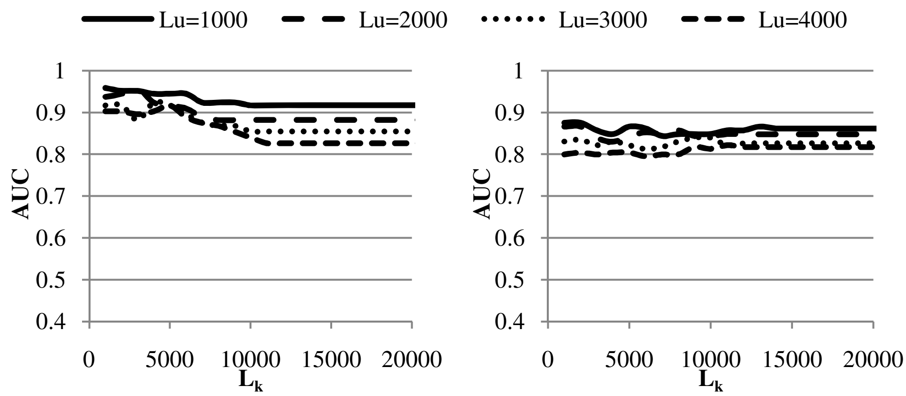

Fig. 5. The performance (AUC) of the proposed verification models on the English part of the training (left) and test (right) corpora with d=d_1, n=5, and different sizes of L_u

Local Settings. To extract local settings, we examine each subset of problems of the training set belonging to a certain language separately. As can be seen in Table 1, the Spanish part of the PAN-2013 training is very limited and it does not enable the

Table 3. F1 scores of the proposed models, the participants and meta-model of PAN-2013

	Overall	English	Greek	Spanish	PAN-2013 rank
Seidman [17]	0.753	0.800	**0.833**	0.600	1st
Halvani [5]	0.718	0.700	0.633	0.840	2nd
Layton et al. [15]	0.671	0.767	0.500	0.760	3rd
Jankowska et al. [7]	0.659	0.733	0.600	0.640	5th
Van Dam [2]	0.600	0.600	0.467	0.760	11th
Meta-model [9]	**0.814**	**0.867**	0.690	**0.898**	-
Our method (global settings)	0.729	0.633	0.767	0.800	-
Our method (local settings)	0.788	0.800	0.767	0.800	-

Table 4. AUC scores of the proposed models, the participants, and meta-model of PAN-2013

	Overall	English	Greek	Spanish	PAN-2013 rank
Jankowska et al. [7]	0.777	0.842	0.711	0.804	1st
Seidman [17]	0.735	0.800	**0.830**	0.600	2nd
Ghaeini [4]	0.729	0.837	0.527	**0.926**	3rd
Meta-model [9]	0.841	0.821	0.756	**0.926**	-
Our method (global settings)	0.789	0.795	0.787	0.917	-
Our method (local settings)	**0.845**	**0.877**	0.787	0.917	-

estimation of appropriate parameter values (i.e., most of the parameter value combinations give perfect results). Therefore, for the Spanish language we used the global parameter settings. Moreover, to enrich the English part of the training corpus, we augmented it by adding more problems based on variations of the initial problems in English. For instance, from a problem with three documents of known authorship we can produce five more problems taking all available subsets of the three known documents as separate verification problems. That way, we formed an augmented version of the English part of the training corpus consisting of 24 problems, all of them variations of the initial 10 problems.

The extracted local parameter settings can be seen in Table 2. The global parameter values coincide with those of the Greek part of the corpus. As already mentioned the global parameter settings were selected for the Spanish part. As concerns English, for parameters n and d, the selected values remain the same with the global settings but L_u and L_k are different (smaller).

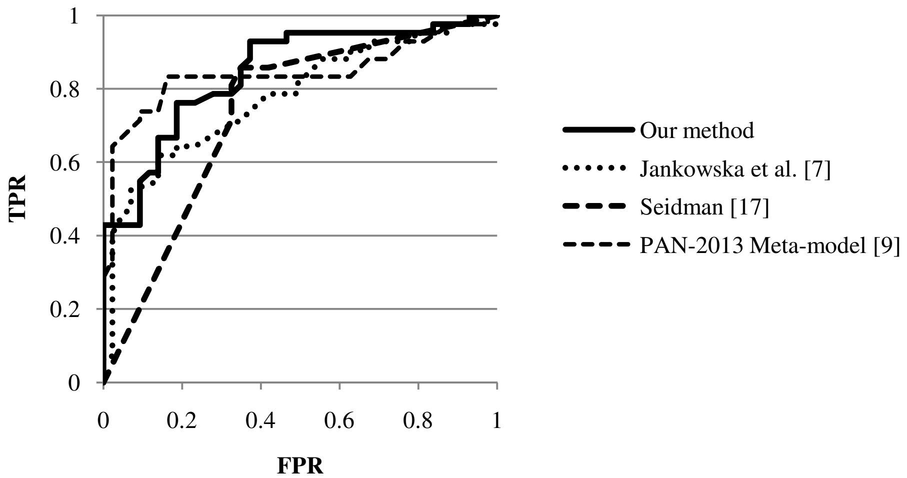

Fig. 6. The ROC curves of the proposed approach (local settings) on the test corpus (all 3 languages) and the corresponding curves of the PAN-2013 winners and the PAN-2013 meta-model

The performance of the verification models on the English part of the training and test corpora for d=d_1, n=5 and different values of L_u and L_k is shown in Figure 5. Apparently, low values of L_u seem to be the most effective ones. This is consistent in both training and test corpora. When compared with the results depicted in Figure 4, we see that the pattern obtained from the entire training corpus (also valid for the Greek part of the corpus) does not apply to the English part of the corpus. For the latter, the most effective option is to extract profiles of equal and small size (1,000 character n-grams) from both the unknown document and the documents of known authorship, that is only the most frequent character n-grams are necessary to achieve good performance.

Comparison with PAN-2013 Participants. As already mentioned, the evaluation procedure we followed is directly comparable with the one performed in the framework of the PAN-2013 competition on authorship verification [9]. Therefore, we can directly compare our results with those of the PAN-2013 participants.

Table 3 shows the performance in terms of F_1 of the binary answers of our method on the test corpus with global and local settings. Overall results as well as language-specific results are presented. Moreover, the corresponding results of the top performing PAN-2013 participants together with the only profile-based participant method [2] and the meta-model combining all submitted methods are reported. The proposed approach based on local settings outperforms every single PAN-2013 participant when the overall performance (F_1) is considered. However, the meta-model continues to be the overall best performing model. On the other hand, our approach achieves more balanced performance in all three languages in comparison to the meta-model. The version of our method based on global settings is also very effective with the exception of the English part and would be ranked 2nd at PAN-2013.

The evaluation of our approach based on the AUC scores of the probability estimates for the entire test corpus and its every language-specific part, is shown in Table 4. Again, the corresponding results for the best performing methods from PAN-2013 and the meta-model combining all submitted models are reported. The proposed approach based on local settings outperforms all others in overall AUC including the PAN-2013 meta-model. Our method achieves the best results in the English part and it is very close to the best results in the Spanish part of the test corpus.

Finally, Figure 6 depicts the ROC curves for the whole test corpus (including 3 languages) of our method (using local settings) and the corresponding curves of the two best performing models at PAN-2013 as well as the meta-model combining all submitted methods. As can be seen, our approach clearly outperforms the methods of [7] and [17]. It also outperforms the meta-model for large FPR values. On the other hand, the meta-model is more effective for low FPR values. This means that when the cost of false positive errors (i.e., incorrect assignment of a document to an author) is considered high, the meta-model wins. In contrast, when the cost of the false negative errors (i.e., miss of a real assignment) is considered high, our approach is better.

5 Conclusions

In this paper, we examined a profile-based method for authorship verification. In contrast to prior evidence, we demonstrated that a profile-based method can be very effective in this task. Our approach is better than any single PAN-2013 participant achieving higher overall F_1 and AUC scores. Moreover, it is very competitive to the heterogeneous meta-model [9], especially when false negatives have high cost.

The fact that the proposed method is less effective in the Greek part of the corpus is partially explained by the difficulty of this corpus since PAN-2013 organizers took special care so that the texts by different authors to be stylistically similar and the texts by the same author to be stylistically dissimilar. This difficulty is reflected in the average performance of PAN-2013 participants on this part of the corpus that was significantly lower with respect to the rest of the corpus. However, a better explanation is that PAN-2013 organizers used a variation of CNG to find stylistically similar or dissimilar texts [9]. It can be claimed, therefore, that the Greek part of the PAN-2013 corpus is negatively biased for approaches based on modifications of CNG.

The proposed approach is fully-trainable. Although the training corpus used in this study is not large, we managed to extract language-specific parameter settings improving the performance in comparison with the case when global settings are used. The performance patterns are consistent in both training and test corpora demonstrating the robustness of our method.

The proposed approach belongs to the family of intrinsic verification methods where no external resources are used by the verification model. Given that extrinsic models seem to be very effective in authorship verification, it could be interesting to investigate how our method can be modified to also use external resources and transform the verification task from a one-class problem to a multi-class problem. Another important future work dimension is to apply the discussed method to verification problems with short documents (e.g., tweets) where the profile-based paradigm has an inherent advantage over instance-based methods.

References

1. Aggarwal, C.C., Zhai, C.X.: A Survey of Text Classification Algorithms. In: Aggarwal, C.C., Zhai, C.X. (eds.) Mining Text Data, pp. 163–222. Springer (2012)
2. van Dam, M.: A Basic Character n-gram Approach to Authorship Verification – Notebook for PAN at CLEF 2013. In: Forner et al (eds.) [3] (2013)
3. Forner, P., Navigli, R., Tufis, D. (eds.): CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
4. Ghaeini, M.R.: Intrinsic Author Identification Using Modified Weighted KNN – Notebook for PAN at CLEF 2013. In: Forner et al (eds.) [3] (2013)
5. Halvani, O., Steinebach, M., Zimmermann, R.: Authorship Verification via k-Nearest Neighbor Estimation – Notebook for PAN at CLEF 2013. In: Forner et al (eds.) [3] (2013)
6. Holmes, D.I.: Authorship attribution. Computers and the Humanities 28, 87–106 (1994)
7. Jankowska, M., Kešelj, V., Milios, E.: Proximity based One-class Classification with Common n-Gram Dissimilarity for Authorship Verification Task – Notebook for PAN at CLEF 2013. In: Forner et al (eds.) [3] (2013)
8. Juola, P.: Authorship Attribution. Foundations and Trends in IR 1, 234–334 (2008)
9. Juola, P., Stamatatos, E.: Overview of the Author Identification Taskat PAN 2013. In Forner et al (eds.) [3] (2013)
10. Keselj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based Author Profiles for Authorship Attribution. In: Proc. of the Pacific Association for Computational Linguistics, pp. 255–264 (2003)
11. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring Differentiability: Unmasking Pseudonymous Authors. Journal of Machine Learning Research 8, 1261–1276 (2007)
12. Koppel, M., Schler, J., Argamon, S.: Authorship Attribution in the Wild. Language Resources and Evaluation 45, 83–94 (2011)
13. Koppel, M., Schler, J., Argamon, S., Winter, Y.: The "Fundamental Problem" of Authorship Attribution. English Studies 93(3), 284–291 (2012)
14. Koppel, M., Winter, Y.: Determining if Two Documents are by the Same Author. Journal of the American Society for Information Science and Technology 65(1), 178–187 (2014)
15. Layton, R., Watters, P., Dazeley, R.: Local n-grams for Author Identification – Notebook for PAN at CLEF 2013. In: Forner et al (eds.) [3] (2013)
16. Sanderson, C., Guenter, S.: Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. In: Proc. of the International Conference on Empirical Methods in Natural Language Engineering, pp. 482–491 (2006)
17. Seidman, S.: Authorship Verification Using the Impostors Method – Notebook for PAN at CLEF 2013. In: Forner et al (ed.) [3] (2013)
18. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics 26(4), 471–495 (2000)
19. Stamatatos. E.: Author Identification Using Imbalanced and Limited Training Texts. In: Proc. of the 4th International Workshop on Text-based Information Retrieval (2007)
20. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60, 538–556 (2009)
21. Stamatatos, E.: Intrinsic Plagiarism Detection Using Character n-gram Profiles. In: Proc. of the 3rd Int. Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (2009)
22. Veenman, C.J., Li, Z.: Authorship Verification with Compression Features – Notebook for PAN at CLEF 2013. In: Forner et al (eds.) [3] (2013)
23. Vilariño, D., Pinto, D., Gómez, H., León, S., Castillo, E.: Lexical-Syntactic and Graph-Based Features for Authorship Verification – Notebook for PAN at CLEF 2013. In: Forner et al (eds.) [3] (2013)