An Improved *Impostors* Method for Authorship Verification

Nektaria Potha and Efstathios Stamatatos

University of the Aegean
83200 - Karlovassi, Greece
{nekpotha,stamatatos}@aegean.gr

Abstract. Authorship verification has gained a lot of attention during the last years mainly due to the focus of PAN@CLEF shared tasks. A verification method called *Impostors*, based on a set of external (impostor) documents and a random subspace ensemble, is one of the most successful approaches. Variations of this method gained top-performing positions in recent PAN evaluation campaigns. In this paper, we propose a modification of the *Impostors* method that focuses on both appropriate selection of impostor documents and enhanced comparison of impostor documents with the documents under investigation. Our approach achieves competitive performance on PAN corpora, outperforming previous versions of the *Impostors* method.

Keywords: Authorship analysis, Authorship verification, Text categorization

1 Introduction

Authorship verification is the task of examining whether two (or more) documents are written by the same author [10, 11, 13]. It is a fundamental task in authorship analysis since any authorship attribution problem can be decomposed into a series of verification problems [9]. In comparison to closed-set attribution, the verification task is more challenging since it focuses on whether the candidate author and the text under investigation have a *similar enough* style rather than what candidate author is *the most similar*. On the other hand, an advantage of verification over closed-set attribution is that the performance of a verification method is affected by less factors since the candidate set size is always singleton and the distribution of training texts over the authors is not so important. Authorship verification methods have been applied in several applications in humanities [18, 7] and forensics [5]. Recently, a series of related PAN@CLEF shared tasks were organized attracting multiple submissions [17, 16].

The *Impostors* method was introduced by Koppel & Winter [11] and so far, it is one of the most successful approaches. Variations of this method won first places in PAN-2013 and PAN-2014 shared tasks in authorship verification [14, 8]. This method uses a set of external documents by other authors (with respect to the ones under investigation) and builds a simple random subspace ensemble. Essentially, it attempts to transform the verification problem from a one-class classification task to a binary classification task

since it calculates whether the texts by the candidate author or the impostors are closer to the disputed texts.

In this paper, we propose a modified version of the *Impostors* method that enhances its performance. Rather than selecting the impostor texts randomly, we propose to use the texts with the highest min-max similarity to the texts under investigation. To use a metaphor, in a police lineup it doesn't make sense to draw the suspects from the general population. Rather, all the suspects should have similar characteristics. Another weakness of the original method is that it disregards cases where at least one impostor text is found more similar to the disputed text in comparison to the texts of the candidate author. To compensate, we propose to rank similarities in decreasing order and take into account the position of the candidate author's text. The proposed approach is evaluated on several PAN corpora and achieves very competitive results.

2 Previous Work

There are two main paradigms in authorship verification. *Intrinsic methods* perform analysis only on the documents under investigation and handle the verification problem as a one-class classification task. They are robust since they do not require external resources and fast since they analyse a few documents [13, 6, 4]. On the other hand, *extrinsic methods* analyse an additional set of external documents and transform the verification problem to a binary classification task [11, 19, 1]. They are usually more effective [17, 16]. From another perspective, a set of verification approaches consider verification problems as instances of a binary classification task and attempt to train a classifier that can distinguish between positive (same-author) and negative (different-author) problems [2, 12]. Such methods heavily depend on the properties of the training corpus.

A modified version of the *Impostors* method, called *General Impostors* (GI) was introduced by Seidman [14]. Since the original method only handles pairs of documents, GI considers the case where multiple documents by the candidate author are available (following the guidelines of PAN shared tasks). Another modification is proposed by Khonji & Iraqi [8]. They focus on the GI weakness of disregarding cases where at least one impostor document is found more similar to the disputed text than the candidate author's text and they utilise the similarity information from those cases. However, the absolute similarity score may significantly differ when different sets of documents are used. Based on that, in this paper we introduce the use of the ranking information. Yet another modification of the *Impostors* method is described by Gutierrez et al. [3]. They propose an aggregate function that iterates over document pairs and applies homotopy-based classification.

3 The Proposed Method

The GI method accepts as input data a set of documents by the same author (known documents) and exactly one document of disputed authorship (unknown document) and provides a score in [0,1] indicating whether the unknown and known documents are by the same author [14]. This score can be viewed as the probability of a positive

answer and can be transformed to a binary answer given an appropriate threshold. GI requires a set of external documents by other authors (with respect to the ones included in the documents under investigation). It randomly selects a subset of these external documents to serve as impostors. Then, it builds a random subspace ensemble by selecting randomly in each repetition a subset of features and a subset of impostors and calculates the similarity of (both known and unknown) documents with impostors [14]. The main idea is that if the known and unknown documents are by the same author, then the known documents will outperform impostors in terms of similarity with the unknown document.

In this paper, we propose a modification of GI (see Algorithm 1) that attempts to improve the following points:

1. *Impostor selection*: Instead of selecting the impostor documents for each verification problem randomly, we propose to select the external documents with the highest min-max similarity [11] score with respect to the known documents. That way, we increase the probability to consider challenging impostor documents that have at thematic or stylistic similarity with the known documents.
2. *Ranking information*: We only compare the impostor document with the unknown document (rather than with both the known and unknown documents). We consider the impostor document as a direct competitor of the known document and therefore we want to know what of these two is more similar to the unknown document. Moreover, instead of taking into account only the cases where the known document is found more similar to the unknown document than all the impostors, we rank (decreasingly) the similarities of both known document and the impostors and consider the ranking position of the known document. For example, if a known document is found to be more similar to the unknown document than all but one impostor across all repetitions (not necessarily the same impostor each time), the original GI method will return a score of 0 while our method will provide a score of 0.5.

4 Experiments

4.1 Setup

To evaluate the proposed approach and compare it with the original GI method and its most important variations, we use the corpora developed at PAN evaluation campaign on authorship verification in 2014 and 2015. These corpora include multiple verification problems in four languages (Dutch, English, Greek, and Spanish) and cover several genres (newspaper articles, essays, reviews, literary texts, etc.) Separate training and evaluation parts are provided for each corpus. In PAN-2014, known and unknown documents within a verification problem have thematic similarities and belong to the same genre. On the other hand, in PAN-2015, known and unknown documents within a problem may belong to different genres and their thematic areas may be distinct which make the task even harder. More details about these corpora as well as evaluation results of PAN participants are provided in [17, 16].

```
Data: D_known, d_unknown, D_external
Parameters: repetitions,|Impostors_problem|,|Impostors_repetition|,rate
Result: FinalScore
for each d_known ∈ D_known do
    for each impostor ∈ D_external do
        MinMax(impostor) = minmaxSimilarity(impostor, d_known);
    end
    Select Impostors_problem ⊂ D_external with highest MinMax(:);
    /* Select Impostors_problem ⊂ D_external randomly                    */
    Set Score(d_known) = 0;
    repeat repetitions times
        Select Impostors_repetition ⊂ Impostors_problem randomly;
        Select rate% of features randomly;
        for each impostor ∈ Impostors_repetition do
            Sim(impostor) = similarity(impostor, d_unknown);
            /* Sim(impostor) =
                similarity(impostor, d_known) * similarity(impostor, d_unknown)   */
        end
        Sim_known = similarity(d_known, d_unknown);
        /* Sim_known = similarity(d_known, d_unknown)^2                  */
        Rank S = Sim(:) ∪ Sim_known in decreasing order;
        pos = position of Sim_known in S;
        Score(d_known) = Score(d_known) + 1/(repetitions * pos);
        /* if Sim_known > max(Sim(:)) then
            Score(d_known) = Score(d_known) + 1/repetitions;
        end
                                                                          */
    end;
    FinalScore = aggregate(Score(:));
end
```

Algorithm 1: The proposed method. Changes with respect to the original GI are shown in blue. Original GI is shown in comments.

The GI method and the proposed variation have several parameters that need to be set. Previous studies attempted to fine-tune these parameters separately [14, 8]. To simplify this process, we focus on fine-tuning parameter $a = |Impostors_{problem}|$ and then use $repetitions$ = a/5 and $|Impostors_{repetition}|$ = a/10. Moreover, we use character 5-grams as features, a fix $rate$ = 0.5 and the *min-max similarity* function. The *aggregate* function is selected among min, max, and average for each training corpus separately. Most of the times, average is selected [14]. Since GI is a stochastic algorithm, each experiment is repeated five times and we report average Area Under the ROC curve (AUROC) measures as used at PAN-2014 and PAN-2015 evaluation campaigns. The set of external documents ($D_{external}$) is constructed for each corpus separately. We submit queries in Bing search engine using significant (with highest *tf-idf*) words from the set of known documents of the training corpus and download the first results. More than 1,000 documents per corpus were downloaded and html tags were stripped off. No further pre-processing is performed.

4.2 Results

For each one of the PAN-2014 and PAN-2015 authorship verification corpora, we report the performance of our implementation of the original GI method and the proposed variation. To study the contribution of each proposed change described in Section 3 separately, we also report performances of taking into account only the impostor selection

Table 1. AUROC results of the proposed approach and other variations of the *Impostors* method.

	PAN14-DE	**PAN14-DR**	**PAN14-EE**	**PAN14-EN**	**PAN14-GR**	**PAN14-SP**	**PAN15-DU**	**PAN15-EN**	**PAN15-GR**	**PAN15-SP**
Khonji & Iraqi (2014)	0.913	**0.736**	0.590	0.750	0.889	**0.898**				
Gutierrez et al. (2015)							0.592	0.739	0.802	0.755
Original GI	0.947	0.660	0.618	0.649	0.772	0.604	0.667	**0.803**	0.656	0.785
Proposed-1	0.970	0.704	0.565	0.738	0.520	0.540	0.662	0.765	0.811	0.825
Proposed-2	0.901	0.698	0.655	0.634	0.860	0.772	0.595	0.786	0.742	0.802
Proposed-full	**0.976**	0.685	**0.762**	**0.767**	**0.929**	0.878	**0.709**	0.798	**0.844**	**0.851**

change (Proposed-1) and only the ranking information change (Proposed-2). Additionally, we include the performance of other variations of the *Impostors* method as described by Khonji & Iraqi [8] and Gutierrez, et al. [3]. The AUROC evaluation results are presented in Table 1. As can be seen, the proposed approach outperforms in all but one case (PAN15-EN) the original GI method, in most of the cases by a large margin. The proposed method is also very competitive with respect to Khonji & Iraqi [8], the overall winner of PAN-2014. There is a mixed picture as concerns the contribution of the impostor selection change and the ranking information change and it is not clear which one of them is most important. However, their combination (proposed-full) is better than each one of them in all but one case (PAN14-DR).

5 Conclusion

Two main changes of the *Impostors* method are proposed in this paper. The first change makes the selection of impostor documents per verification problem a deterministic procedure ensuring that impostors will have similar characteristics with the candidate author's texts. The second change attempts to enrich the information that is kept in each repetition of the random subspace ensemble. Experiments in several authorship verification corpora demonstrate that the combination of these changes significantly enhance the performance and the proposed approach is competitive, if not better, than another variation of GI that won the first place in PAN-2014 evaluation campaign [8]. The presented results further attest the effectiveness of the *Impostors* method and future work can more thoroughly examine the use of alternative text representation schemes and the profile-based paradigm [15].

References

1. Bagnall, D.: Author Identification using multi-headed Recurrent Neural Networks. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
2. Fréry, J., Largeron, C., Juganaru-Mathieu, M.: UJM at clef in author identification. In: CLEF 2014 Labs and Workshops, Notebook Papers. CLEF and CEUR-WS.org (2014)

3. Gutierrez, J., Casillas, J., Ledesma, P., Fuentes, G., Meza, I.: Homotopy Based Classification for Author Verification Task—Notebook for PAN at CLEF 2015. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
4. Halvani, O., Winter, C., Pflug, A.: Authorship verification for different languages, genres and topics. Digital Investigation 16, S33 – S43 (2016)
5. Iqbal, F., Khan, L.A., Fung, B.C.M., Debbabi, M.: e-mail authorship verification for forensic investigation. In: Proceedings of the 2010 ACM Symposium on Applied Computing. pp. 1591–1598. ACM (2010)
6. Jankowska, M., Milios, E.E., Keselj, V.: Author verification using common n-gram profiles of text documents. In: Proceedings of COLING, 25th International Conference on Computational Linguistics. pp. 387–397 (2014)
7. Kestemont, M., Stover, J.A., Koppel, M., Karsdorp, F., Daelemans, W.: Authenticating the writings of julius caesar. Expert Systems with Applications 63, 86–96 (2016)
8. Khonji, M., Iraqi, Y.: 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 (2014)
9. Koppel, M., Schler, J., Argamon, S., Winter, Y.: The fundamental problem of authorship attribution. English Studies 93(3), 284–291 (2012)
10. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, 1261–1276 (2007)
11. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the American Society for Information Science and Technology 65(1), 178–187 (2014)
12. Pacheco, M., Fernandes, K., Porco, A.: Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
13. Potha, N., Stamatatos, E.: 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 (2014)
14. Seidman, S.: Authorship Verification Using the Impostors Method. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
15. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60, 538–556 (2009)
16. Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., Stein, B.: Overview of the author identification task at PAN 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum (2015)
17. Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., Barrón-Cedeño, A.: Overview of the author identification task at PAN 2014. In: Working Notes for CLEF 2014 Conference. pp. 877–897 (2014)
18. Stover, J.A., Winter, Y., Koppel, M., Kestemont, M.: 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 (2016)
19. Veenman, C., Li, Z.: Authorship Verification with Compression Features. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)