# An Analysis Framework for Hybrid Authorship Verification

Seifeddine Mechti [1], Maher Jaoua[2] , Rim Faiz[1,3], and Lamia Hadrich Belguith[2]

[1]LARODEC Laboratory, ISG of Tunis B.P.1088, 2000 Le Bardo, Tunisia
mechtiseif@gmail.com, Rim.faiz@ihec.rnu.tn
[2]ANLP Group, MIRACL Laboratory, University of Sfax, BP 1088, 3018, Sfax Tunsia
[3]IHEC of Carthage, 2016 Carthage Présidence, Tunisia
{maher.jaoua,l.belguith}@fsegs.rnu.tn

**Abstract.** Given a set of candidate authors for whom some texts of undisputed authorship exist, attribute texts of unknown authorship to one of the candidates is called Author verification. This problem acquired great attention due to its new applications in forensic analysis, e-commerce and plagiarism detection. The author verification task is of great help in the plagiarism detection process. Indeed, the probability of plagiarism increases where two parts of a document are not assigned to the same author. This paper introduces an analysis framework for hybrid authorship verification. In fact, the proposed method takes advantage of a large set of linguistic features to fully address the identification of the document's author. These features are explored to build a machine-learning process. We obtained promising results by relying on PAN@CLEF 2014 English literature corpus.

## 1 Introduction and Related Works

Although the writing style analysis is an old research area and has been applied successfully to solve many problems, notably authorship attribution, it is obvious that its application to identify the authors of anonymous texts still needs to be investigated further.

Author attribution consists in identifying the author, one of a list, who wrote a particular anonymous text, this categorization focus on open-set[1] or closed-set[2] classification problems [1]. A more difficult author attribution task is the author verification. In this task, we addresses a non-factoid question: "was a particular text written by a well-defined author?".

Recently, the issue of determining the authorship of a document acquired great attention due to its new applications in forensic analysis, plagiarism detection, forensic linguistics, and e-commerce. Additionally, the author identification task is of great

[1] The true author of the disputed text is not necessarily included in the set of candidate authors.
[2] The true author of the disputed text is necessarily included in the set of candidate authors.

help in the plagiarism detection process. Indeed, the probability of plagiarism increases where two parts of a document are not assigned to the same author. Forensic analysis or the analysis of the paternity of documents for legal purposes can contribute to several investigations focusing on various linguistic characteristics.

We grouped methods of authors identification essentially into three categories. The first one is based on a linguistic analysis. The second method is based on various statistical analyses. The more recent third one uses machine learning algorithms.

The basic idea of the stylistic methods is based on the modeling of authors from a linguistic point of view. We cite as an example the works of Li et al. who have focused on topographic signs [2] as well as the works of Zheng et al. who were interested in the co-occurrence of character n-grams [3]. Other works were concerned with the distribution of function words [4] or the lexical features [5]. In another work, Raghavan et al. exploited grammars excluding the probabilistic context to model the grammar used by an author [6]. Feng et al. based their work on the syntactic functions of words and their relationships in order to discern entity coherence [7]. Other studies have focused on the semantic dependency between the words of written texts by means of taxonomies and thesaurus [8].

The first attempts emerged in the studies of [10], constituting the first real great statistical study of texts; they compared the occurrence frequency of words such as verbs, nouns, determinants, prepositions, conjunctions, and pronouns.

In the last few years, a number of new methods which are based on various statistical tools have been presented in order to discriminate between the potential authors of a text. Among these methods, we find inter-textual distance [11], the Delta method [12], the LDA distribution [13] and the KL divergence distance [14].

Recently, from a machine learning point of view [1], author verification method is intrinsic or extrinsic, intrinsic methods use only the known texts and the unknown text of the problem[3] and extrinsic methods uses external documents of other authors for each problem.

The training corpuses are represented in a varied form; we can consider each text as a vector in a space with several variables. In addition, a variety of powerful algorithms can be used to build a classification model, including discriminating analysis [15], SVM [16], decision trees [17], the neural network [4], genetic algorithms [18]. [17] adopt a machine learning approach based on several representations of the texts and on optimized decision trees which have as entry various attributes. This method obtains the first rank in competitive conference Pan@clef2014 only in English essays [1].

The rest of the paper is organized as follows. Section 2 presents the proposed method. Section 3 provides the implementation of the HaiTay System. Section 4 presents experimentations and evaluation. Finally, Section 5 draws our conclusions.

## 2 Proposed Method

Hybridization has always been considered an interesting track because it overcomes the limitations of combined approaches. It is with this objective in mind that we tried

[3] We call "problem" any test document whose paternity is unknown.

to experiment with learning techniques on all the stylistic and statistical features that have shown their efficiency in the literature. The basic idea is to create for each text T, whose belonging to an author A we want to verify, a sub corpus which includes all the texts written by this author and the texts that are close to it in terms of distance. Thus, if the text was written by author A then there is a high probability that we recognize the style via the stylistic and statistical features of author A's texts belonging to the corpus. On the other hand, if A is not the writer of T then there is a good chance that it is assigned to another author selected from the rest of the sub corpus.

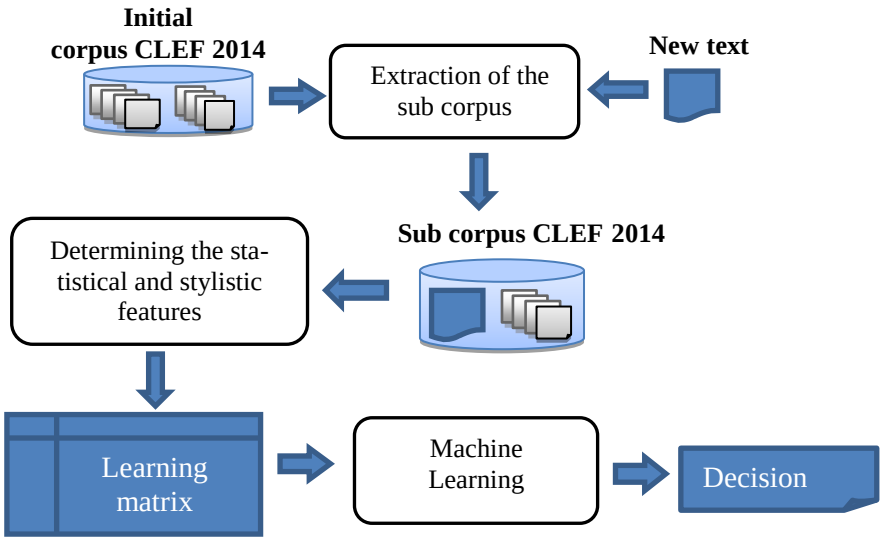


**Fig. 1.** Steps of the proposed method

## 3 Implementation of the HyTAI System

In order to implement the proposed method, we developed a system called HyTAI (Hybrid Tool for Author Identification) whose modular decomposition follows the proposed method. Thus, we used the Delta rule in the extraction module of the sub corpus to calculate the distance between two texts. Also, we used the OpenNLP for the extraction of the stylistic and statistical features.

To calculate the distance between two documents, we used the Delta distance proposed by Burrows et al. (Burrows 2002). This distance, which takes into account the most frequent words, is characterized by the following formula:

$$\Delta(Q, A_j) = 1/m \sum_{i=1}^{m} |Zscore(t_{iq}) - Zscore(t_{ij})|$$

where

$$Z\ score\ (t_{tij}) = \frac{tfr_{ij} - mean_i}{sd_i}.$$

Note that $tfr_{ij}$ is the frequency of the term $t_i$ in the document $Dj$ while $mean_i$ is the mean and $sd_i$ is the standard deviation.

It should be noted that if two texts are quite close, then delta tends toward 0. Similarly, the value m may vary from one corpus to another and that is why we conducted an experiment to have the value determined (see next section). For the training sub corpus, we choose the nearest texts of a document to be checked in such a way that a balance is achieved between the texts written by the author to be identified and the texts that do not belong to that author.

In order to extract the stylistic and statistical features, we used tools from the Apache OpenNLP library, which contains a set of functions that can segment texts and perform the syntactic and lexical analyses. We calculated the frequency of lexical features, the ratio V / N – where V is the hapax's size and N is the text length – and the average length of sentences. Regarding parsing, also conducted through the OpenNLP, we extract the number of nouns, the number of verbs, the number of adjectives, the number of adverbs and the number of prepositions.

Then to extract the features related to the model of the language, we consider the text as a simple sequence of characters and determine the frequencies of the letters, the punctuation marks and the numeric characters as well as n-grams.

## 4 Results Analysis

In this section, we present the experimental results of our method for the identification of authors. We first describe the corpus and the measures of evaluation. Then, we present the performance of our system in the identification of anonymous authors.

**Table 1.** Dataset description

| Number of proposed problems | Number of known documents / author | Average length of unknown documents | Average length of knowns documents |
|---|---|---|---|
| 200 problems | 2.65 documents | 806.86 words | 845.30 words |

The dataset includes a set of folders from the PAN@CLEF 2014 computational conference. Each folder includes up to five documents and a test document in English. The length of the documents varies from a few hundred to a few thousand words. We should note that we carried out the experiment with the 200 existing problems in the corpus.

In our evaluation, we compare different variants of our proposed stylistic, Statistical n-grams and hybrid author verification methods:

- Stylistic method using lexicals (**le**), syntactic (**sy**), characters (**ch**) and stylistic (**st** = **le**+ **sy** + **ch**) features.
- Statistical method using the Delta rule (**Statis**).

- Machine learning method using SVM, decision table, decision trees, naïve bayes, etc.
- A hybrid method, based on SVM, using both the categories of stylistic features and the Delta rules (**St**+ **Statis** + **Ma**) as described in Ffigure 6;
- A baseline method using n-grams with n = 3, 4, 5, 6 and 7.

The evaluation score c@1[19] has the advantage of taking into account the documents that the classifier is unable to assign to a category. For each problem, each score greater than 0.5 is considered as a positive response and the document is indeed the property of the author in question. Each score below 0.5 is considered as a negative response and therefore the test document does not belong to this author. Nevertheless, all the scores equal to 0.5 correspond to the outstanding problems where the answer will be "I don't know". Then, c@1 is defined as follows:

$$c@1 = (1/n)*(nc + (nu*nc/n)) \quad [19] \tag{3}$$

where n is the number of problems; nc is the number of correct answers; nu is the number of unanswered problems.

The histograms below present the experiments we conducted to obtain the best possible documents paternity.

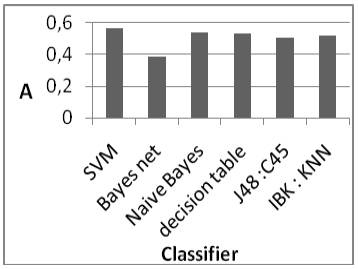


**Fig. 2.** The accuracy of different classifiers

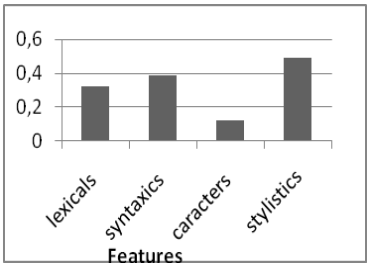


**Fig. 3.** The c@1 performance of different types of features

Figure 2 shows the accuracy reached with a test set of six well-known classifiers in order to select the best. This is determined with all the stylistic features and the n-gram features (variation of n between 3 and 7). The best accuracy has been achieved by the use of the SVM algorithm with a slight advantage compared to the Naïve Bayes classifier.

Using the SVM classifier, we examine the three categories of features, each category apart and then the 3 gathered categories.

The result presented in Figure 3 shows that the character features are not very powerful in determining the authors of documents whose origin is unknown. On the other hand, the syntactic features give encouraging results. The combination of these features provides a better performance than the use of each feature alone.

Figure 4 depicts the c@1 histogram of the n-grams method. This figure shows that accuracy reaches a maximum for n= 3 and 4, and then it decreases with the increase of n. Therefore, the n-gram models reach a good performance between 3 and 4, and

then they will not be effective. Then, we use the most frequent number of m words between 100 and 400.

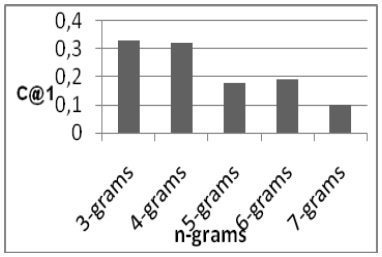


**Fig. 4.** The c@1 performance according to the n-grams methods

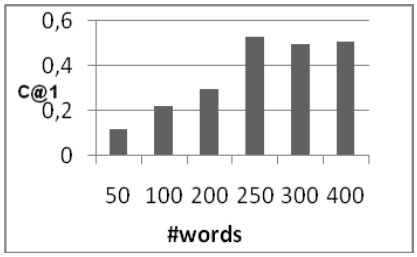


**Fig. 5.** The c@1 performance according to the number of words

Figure 5 shows that the best c@1 measure is obtained based on the SVM algorithm with 250 words. Then it decreases with the increase of number of words.

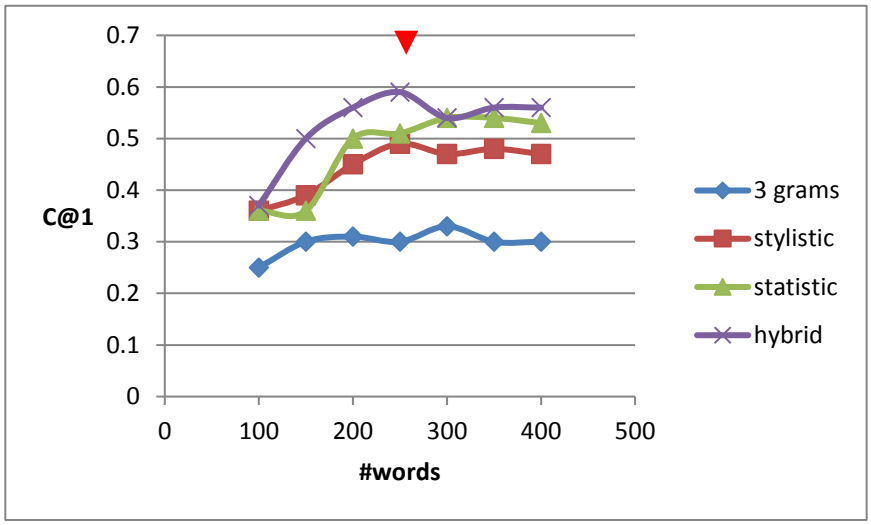


**Fig. 6.** The C@1 Performance of different features according to words number

Figure 6 shows that the combination of the syntactic features, the lexical ones and the 3-grams brings an encouraging result in a machine learning process. However, the use of delta method for the classification of documents gave better results than stylistic method, we obtain 0.54 c@1 score.

In the hybrid evaluation set up, this result is somewhat improved by using the Delta method. These measures reach a very good value with the choice of the most frequent 250 words. Our system has proven its effectiveness when the statistical and the stylistic analysis are combined. We have been able to find the unknown author of a document in 59% of cases.

Based on Table 2, we compared the performance of our method with those of the winner of PAN@CLEF 2014 competitive conference for the English essays.

**Table 2.** Performances of our method in comparison with Frery et al.

| | Our method | Frery et al. [17] |
|---|---|---|
| C@1 | 0.59 | 0.71 |
| AUC[4] | 0.6 | 0.72 |

Our classification, compared with the best systems, is encouraging, which shows the effectiveness of our method. With C@1 equal to 0.59 we obtain the 4th Rank.

## 5 Conclusion

In this study, we built a hybrid method by combining linguistic features and n-grams. Through experiments relying on a real-world corpus, we showed that the hybrid method outperforms some other methods since we combine syntactic features, lexical features, n-grams and character features. This demonstrates the great potential of heterogeneous models in detection of document's paternity.

The experiments described in this paper were performed on Pan@CLEF 2014 corpora comprising documents in English. We obtained comparable results to the best performing systems

Our method best configuration is 3 as the n-grams length, only 250 as the number of terms and SVM as the learning algorithm.

As future work, we seek to improve our method using a text-extraction tool. We aim to introduce the idea that the style of the author resides in one part of the document rather than in others.

## References

1. Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast M., Stein, B., Juola P., Sanchez-Perez, M., Barrón-Cedeño, 2014. A.: Overview of the Author Identification Task at CLEF. England
2. LI, J., Zheng, R., Chen, H. 2006. From fingerprint to writeprint. Communication ACM 49(4), 76-82.
3. Zheng, R., Li, J., Chen, H., and Huang, Z. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology, 57(3), 378-393.
4. Vartapetiance, A., Gillam, L.2014. A Trinity of Trials: Surrey's 2014 Attempts at Author Verification. Proceedings of PAN@CLEF.
5. Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., Levitan, S.:2007. Stylistic text classification using functional lexical features Journal of American society of information science and technology 58(6), 802–822.

[4] Area under the roc curves [1]

6. Raghavan, S., Kovashka, A., Mooney, R.2010. Authorship attribution using probabilistic context-free grammars .Proceedings of ACL'10, 38–42.
7. Feng, V.W., Hirst, G..2013. Authorship verification with entity coherence and other rich linguistic features. Proceedings of CLEF'13.
8. Mccarthy, P.M., Lewis, G.A., Dufty, D.F., Mcnamara D.S.. 2006. Analyzing writing styles with coh-metrix.Proceedings of FLAIRS'06, 764-769.
9. Baayen, R.H. 2008.Analyzing Linguistic Data.: A Practical Introduction to Statistics using R".Cambridge, Cambridge University Press, Cambridge.
10. Mosteller, F., Wallace, D. 1964. inference in an Authorship Problem, In Journal of the American Statistical Association, Volume 58, Issue 302, 275-309.
11. Labbé, C. 2003. Inter-Textual Distance and Authorship Attribution. Corneille and Molière, In: Journal of Quantitative Linguistics, 213-231.
12. Burrows, J.: Delta: 2002. A Measure of Stylistic Difference and a Guide to Likely Authorship, In Journal Lit Linguist Computing.
13. Blei, D.M., Jordan, M.I. 2004 Variational methods for the Dirichlet process. In Proceedings of the twenty-first international conference on Machine learning ACM.
14. Hershey, J.R., Olsen P.A., Rennie, S.J. . 2007. Variational Kullback-Leibler divergence for Hidden Markov models. IEEE Workshop on Automatic Speech Recognition and Understanding.
15. Stamatatos, E., Fakotakis, N., Kokkinakis, G.. 2000. Automatic text categorization in terms of genre and author, Computational Linguistics, Volume 26,.471-495.
16. Lee, C., Mani, I., Verhagen, M., Wellner, B., Pustejovsky, J.: 2006.Machine learning of temporal relations". In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. 753-760.
17. Frery J, Largeron ch, and Juganaru-Mathieu, M. 2014. UJM at CLEF in Author Identification. PAN@CLEF2014.England.
18. Moreau, E., Jayapal, A., Vogel, C. 2014. Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm. Notebook for PAN at CLEF 2014. England.
19. Peñas, A. and Rodrigo, A. 2011. A Simple Measure to Assess Non response. In Proc. Of the 49th Annual Meeting of the Association for Computational Linguistics, Vol.1, 1415-1424.