An Empirical Method Exploring a Large Set of Features for Authorship Identification

Seifeddine Mechti
LARODEC Laboratory
ISG of Tunis B.P.1088, 2000 Le Bardo, Tunisia
mechtiseif@gmail.com

Rim Faiz
LARODEC Laboratory
IHEC Carthage, Tunisia
Rim.faiz@ihec.rnu.tn

Maher Jaoua
MIRACL Laboratory
FSEGS, BP 1088, 3018 Sfax, Tunisia
maher.jaoua@fsegs.rnu.tn

Lamia Hadrich Belguith
MIRACL Laboratory
FSEGS, BP 1088, 3018 Sfax, Tunisia
l.belguith@fsegs.rnu.tn

Abstract

In this paper, we deal with the author identification issues of the document whose origin is unknown. To overcome these problems, we propose a new hybrid approach combining the statistical and stylistic analysis. Our introduced method is based on determining the lexical and syntactic features of the written text in order to identify the author of the document. These features are explored to build a machine learning process. We obtained promising results by relying on PAN@CLEF2014 English literature corpus. The experimental results are comparable to those obtained by the best state of the art methods.

1 Introduction

Recently, much more interest has been given to a document authorship because of its application in many domains, such as e-commerce, forensic linguistics, etc. For instance, in the latter, author identification can make many investigations easier. Addi-tionally, the author identification task is very useful 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 task is planned in PAN@CLEF 2016.In addition, forensic analysis or that of the documents paternity for legal purposes can contribute to several investigations focusing on various linguistic characteristics. In the literature, the automation of the author identification task can draw on stylistic or statistical attributes. Currently, machine learning techniques have been used to infer attributes discriminating the authors styles. In this context, we propose a hybrid method combining the stylistic and statistical attributes while relying on measure-ments of inter-textual distances. In this paper, we present the results of our experiments, using several learning techniques. The objective of the work proposed in (Stamatatos et al., 2014) is to determine from a specific list the au-thor who wrote a given text. Thus, for this identification, we should focus on open-set or closed-set classification problems. In this context, we address a non-factoid question: was a particular text written by a well-defined author. This paper is organized as follows: In section 2, we depict the author identification approaches proposed in literature. After that, we present our hybrid method based on the statistical and stylistic analysis. In section 3, we describe the machine learning process. The fourth section shows the experiments carried out together with the sever-al applied tests and algorithms. Then, we compare our simulation results with those obtained by using other methods. Finally, we end up this paper by some concluding remarks, and we propose future research study.

2 Related Work

In this section, we introduce author identification methods classified essentially into three categories. The first one is based on a stylistic analysis. The second class contains techniques relying on various statistical analyses. The third category, which includes more recent methods, uses machine learning algorithms. The basic idea of the stylistic methods is the modeling of authors from a linguistic point of view. For instance, we can mention the works of Li et al.(2006), who focused on topographic signs (Li et al., 2006), as well as the studies of Zheng et al. interested in the co-occurrence of character n-grams (Zheng et al., 2006). Other researchers were concerned with the distribution of function words (Vartapetience et al., 2014) or the lexical features (Argamon et al., 2007). In another work, Raghavan et al.2006 exploited grammars excluding the probabilistic context to model the grammar used by an author (Raghafan et al., 2010). Feng et al. dealt with the syntactic functions of words and their relationships in order to discern entity coherence (Feng et al, 2013). Other surveys studied the semantic dependency between the words of written texts by means of taxonomies and thesaurus (Maccarthy et al, 2006). Concerning statistical methods, the first attempts emerged in (Mostler et Wallace., 1964). They compared the occurrence frequency of words, such as verbs, nouns, articles, prepositions, conjunctions, and pronouns. In the last few years, new methods, based on various statistical tools, have been introduced in order to discriminate between the potential authors of a text. Among these methods, we can mention intertextual distance (Labbé,2014), the Delta method (Savoy, 2013), the LDA distribution (Blei et al, 2004) and the KL divergence distance (Herchey et al., 2007). Indeed, (Labbé,2003) Labb demonstrated the effectiveness of intertextual distance in quantifying the proximity between several texts through a normalized index. Later, he revealed the considerable Corneille contribution in plays written by Moliere . In (Savoy, 2014), Buroows proposed the Delta method in order to identify the unknown documents author. He has suggested selecting 40 to 150 most frequently used words, especially the functional words, while ignoring the punctuation signs. On the other hand, in (Grieve, 2007), researchers demonstrated that the Delta method could offer the best results. To determine the document paternity, the authors introduced a probabilistic model for author identification by addressing several topics (Savoy, 2012). At this level, each corpus is modeled as a distribution of different themes; each theme represents a specific distribution of words. From a machine learning point of view (Stamatatos et al, 2014), author verification method can be either intrinsic or extrinsic. In fact, intrinsic methods use both the known and unknown texts of the problem , while extrinsic methods utilize external documents of other authors for each problem. The training corpuses are represented in different forms. each text is considered as a vector in a space with several variables. In addition, a variety of powerful algorithms, including discriminating analysis (Stamatatos et al,2000), SVM (Lee et al., 2006), decision trees (Zhao et Zobel, 2006), the neural network (Argamon et al., 2007) and genetic algorithms (Moreau et al., 2014), can be used to construct a classification model. Finally, in a critical study carried out by Baayen, the latter showed that the stylistic methods revealed low performances for short texts (Baayen et al, 2008). He also proved that style can change over time or according to the literary genre of the texts (poetry, novels, plays ...). Besides, despite their interesting results, the statistical analysis ignores the writers style. In this case, neither the vocabulary nor the theme of the suspect document is taken into account. Olson criticized some studies which convert the language into mathematical equations (Herchey el al., 2007). We choose hybridization to take advantage of both the stylistic methods and statistics. On the one hand, we use the lexical and syntactic analysis to address the problem of mathematical representation of a text (Section 3.1). On the other hand we apply the Delta rule to gather the writers who have almost the same style (section 3.2).

3 The Proposed Method

The following section describes our hybrid extrinsic method for tauthor identification. First, we will extract the different types of stylistic features (syntactic, lexical and characters) and then the n-grams. In the second step of the authors selection, we will focus on the delta method. The third step will be reserved for the application of the learning model.

3.1 Feature Extraction

In order to extract features, also called style markers, we use the tools of the Apache Open Library . These robust tools allow segmenting the texts and analyzing the necessary syntax and semantics. For the lexical features, obtained by frequency calculations, the text is regarded as a set of tokens. We distinguish between the number of words that appear only once, the ratio V/N (V is the size of the hapaxes , and N is the length of the text), the average sentence length and the number of words which appear twice. Then, we extract the lexical features, such as the number of nouns, verbs, adjectives, adverbs and prepositions. In features extraction, we consider the text as a simple sequence of characters. We also take into account the information concerning the frequencies of letters, punctuation marks (number of colons, exclamation marks, question marks and commas), uppercase and lowercase characters as well as the numerical and alphabetical characters. Finally, we resort to the n-grams classes. We make n vary from 3 up to 7 characters. In fact, a small n=3 and a large one are respectively used to capture the syllables and the punctuation marks and to produce the words.

3.2 Authors Selection

In this step, we select authors in order to prepare the machine learning process. We apply the Delta method on the candidate document and all authors existing documents. For each unknown author, we select the three authors who have the lowest Delta measure with the candidate document. We note that different verification problems (different folders) may share documents of the same authors. For example, the known document of folder EN001 and that of folder EN002 may be written by the same author. Then, we calculate the distance based on the standardized frequencies (Z-score) between two documents Q and A using the following equation:
$D(Q, A_j) = \frac{1}{M}\sum_{i=1}^{m})[Zscore(t_{iq}) - Zscore(t_{ij})]$

Where
$Zscore(t_{ij}) = \frac{tfr_{ij} - mean(i)}{sd(i)}$

tfr_{ij} is the frequency of the term ti in the document Dj, mean represents the average, and sdi denotes the standard deviation. Finally, we use the number of the most common terms between 100 and 400 words.

3.3 Application of a Classification Model

We perform the machine learning process based on the documents of the candidate author and those of the three already selected authors. We use the Weka tool in order to represent the known author and the other three authors by an ARFF file with the already extracted features. In addition, we apply a learning algorithm on this File in order to get a prediction model where the known texts are the positive examples, and documents written by other authors represent the negative examples. This algorithm is determined after applying a test on multiple classifiers, such as: SVM, decision trees, Naive Bayes, decision table and KNN. We choose the algorithm that gives the best performance.

4 Basic characteristichs of our Hybrid method

Hybridization has always been considered as an interesting track because it overcomes the limitations of the combined approaches. The following table 1 presents a comparison between the different methods of author identfication: Verification Model: The intrinsic models use the texts within a verification problem (Zheng et al.,2006), (Feng et al.,2013), (Mostler et Wallace.,1964). In other studies (Labbé, 2014), (Savoy, 2012) Labb and Savoy consider other texts written by different authors and attempt to transform the verification task into a binary classification problem. However, According to PAN@CLEF 2014 and PAN@CLEF 2015, extrinsic models give better results than intrinsic ones (Stamatatos et al.,2014). Classifcation: There are two methods of classification: eager methods, using a supervised learning (Zheng et al.,2006), (Feng et al,2013), and lazy methods that do not apply any algorithm (Mostler et Wallace, 1964), (Labbé, 2003), (Savoy, 2012). In this paper, we resort to supervised learning using SVM. Attribution Paradigm: There are two attribution paradigms (Stamatatos et al, 2000). In the instance based representation each document is represented separately (Feng et al., 2013), (Labb, 2003), (Savoy, 2012). However, the profile based paradigm tries to construct an author profile using all texts of the corresponding author. (Author profile) (Zheng et al.,2006), (Mostler et wallace, 1964). Indeed, we choose the hybrid of the two paradigms, a representation for each document which are then combined in a single author profile. Text analysis: Most of the proposed studies used the part of speech POS tagging (Zheng et al., 2006), (Mostler et wallace, 1964) because of the availability of taggers. Some other studies resorted to intertextual distance (Labb, 2003), (Savoy, 2012). However, our method combines statistical and stylistic features (sections 3.1, 3.2). The following section describes our hybrid extrinsic method for tauthor identification. First, we will extract the different types of stylistic features

Table 1: Author identifcation methods

Author (s)	Verification Model	Classification	Attribution paradigms	text analysis
(Zheng., 2006)	extrinsic	Eager	Author profile	POS taggig
(Feng, 2013)	extrinsic	Eager	Instance based	POS taggig
(Wallace et al., 2011)	extrinsic	Lazy	Author profile	POS tagging
(Labbé ,2014)	intrinsic	Lazy	Instance based	Intertextual distance
(Savoy et al., 2013)	intrinsic	Lazy	Instance based	Delta method
Our Method	extrinsic	eager	Hybrid	Delta metod + POS taggig

(syntactic, lexical and characters) and then the n-grams. In the second step of the authors selection, we will focus on the delta method. The third step will be reserved for the application of the learning model.

5 Experiments and Evaluation

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

5.1 Corpus

The training corpus includes a set of folders from the PAN@CLEF 2014 computational conference. Each folder contains up to five machine learning documents and a test document in English. The length of the documents varies from a few hundred to a few thousand words. It is worth noting that the experiments were carried with the 200 existing problems in the corpus.

5.2 Performance Measures

To assess our results, we adopt the the C@1 measure (Penas et Rodrigo., 2011) AUC and Recall metrics.

Recall

In the context of classification tasks, the terms true positives, true negatives, false positives and false negatives are used to compare the given classification of an item :

TN / True Negative: case was negative and predicted negative
TP / True Positive: case was positive and predicted positive
FN / False Negative: case was positive but predicted negative
FP / False Positive: case was negative but predicted positive

Recall= VP/(VP+FN)

C@1 score

The evaluation score C@1 has the advantage of considering 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, while that below 0.5 is viewed as a negative response. 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 dont know ”. Then, c @ 1 is defined as follows:

c@1 = (1/n)*(nc+(nu*nc/n))
(Penas et Rodrigo, 2011)

where:
n = number of problems ;
nc = number of correct answers ;
nu = number of unanswered problems

AUC score

The AUC is a common evaluation metric for binary classification problems.

the figure 1 present an exmample of AUC plot. Consider a plot of the true positive rate vs the false positive rate as the threshold value for classifying an item as 0 or is increased from 0 to 1: if the classifier is very good, the true positive rate will increase quickly and the area under the curve will be close to 1. If the classifier is no better than random guessing, the true positive rate will increase linearly with the false positive rate and the area under the curve will be around 0.5.

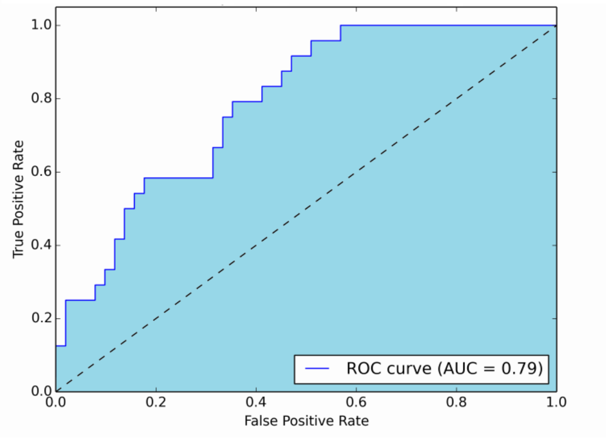


Figure 1: Example of AUC plot

5.3 Result Analysis

The histograms below reveal the experiments conducted to obtain the best possible documents paternity:

Figure 2 (a) shows the accuracy reached with a test set of six well known classifiers in order to select the best one. This accuracy 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 vis-a-vis the Nave Bayes classifier. Figure1 (b) show 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. Combining these features provides better performance than the use of each feature separately. Figure 1(c) depicts the c@1 histogram of the n-grams method. It highlights that accuracy reaches a maximum for n= 3 and 4. Then, it decreases with the increase of n.. After that, we use the most frequent numbers of m words (between 100 and 400). Figure 1(d) shows that the best c@1 measure is given based on the SVM algorithm with 250 words. This measure decreases with the increase of words number.

Figure 3 demonstrates that combining the syntactic features, the lexical ones and the 3 grams brings encouraging results in a machine learning process. However, the use of the Delta method to classify documents gives better results than the stylistic method by which we obtain 0.54 c@1 score. In the hybrid evaluation step, this result is somewhat improved by using the Delta method during the step of authors selection. These measures reach high value with the choice of the most frequent 250 words. Our system has proven its effectiveness when the statistical and the stylistic analysis were combined. Thus, we were able to find the unknown author of a document in 59% of the studied cases. In Table 2, we compare the performance of our method with those of the winner of PAN@CLEF 2014 competitive conference for the English essays. From table 2, we notice that our method is useful in terms of recall. It noticeably outperforms Frery et al.(2014), although C@1 and AUC still need to be further improved. Based on PAN@CLEF 2014 competitive conference (Stamatatos et al, 2014), our classification results are so encouraging, which shows the effectiveness of our method. Focusing on the step of selecting the attributes, we are trying to improve our results in our future work.

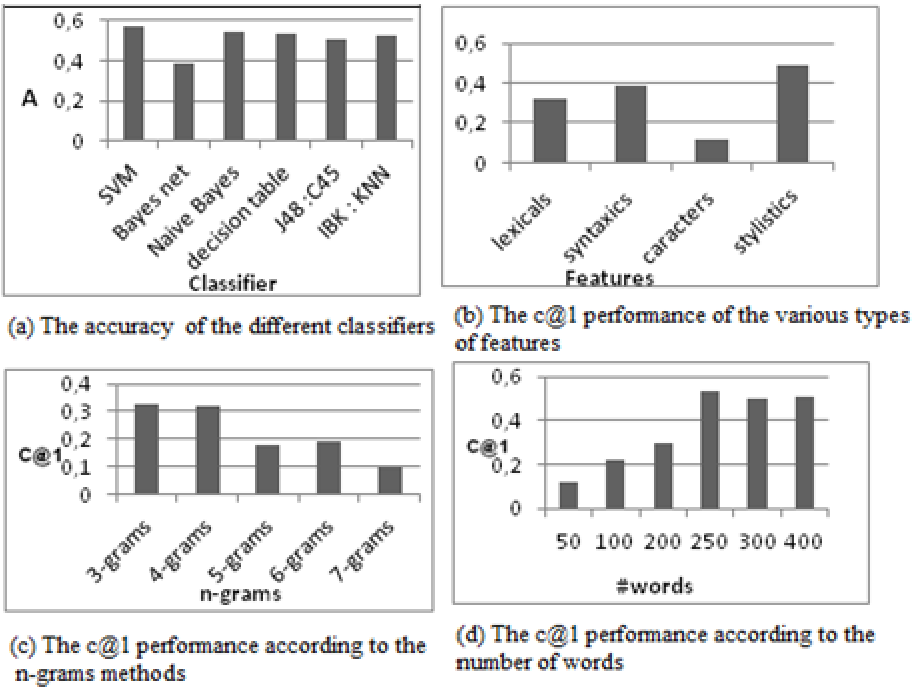
Table 2: Comparison between our performances and Frery el. 2014

	Baseline	Our method	Frery et al.(2014)
C@1	0.53	0.68	0.71
Recall	0.5	0.74	0.72
AUC	0.54	0.6	0.72

6 Conclusion

In this paper, we have focused on author identification problem by applying a machine learning process. Indeed, the introduced hybrid method is essentially based on using both stylistic and statistical characteristics. The experimental results reveal the efficiency of the proposed technique in which we use the Delta method prior to syntactic and lexical features as well as n-grams and character features. We have also proven through the carried experiments how the heterogeneous models allowed us to detect appropriately the document paternity. In future research study, we will try to make our technique more effective by utilizing text extraction tool. The main objective will be to show that the authors style is clear in some specific parts of the written text.

We are also planning to apply our approach on German, Spanish and Greek corpora to show the efficiency of our method in multilingual context.



(a) The accuracy of the different classifiers

(b) The c@1 performance of the various types of features

(c) The c@1 performance according to the n-grams methods

(d) The c@1 performance according to the number of words

Figure 2: author identification histograms

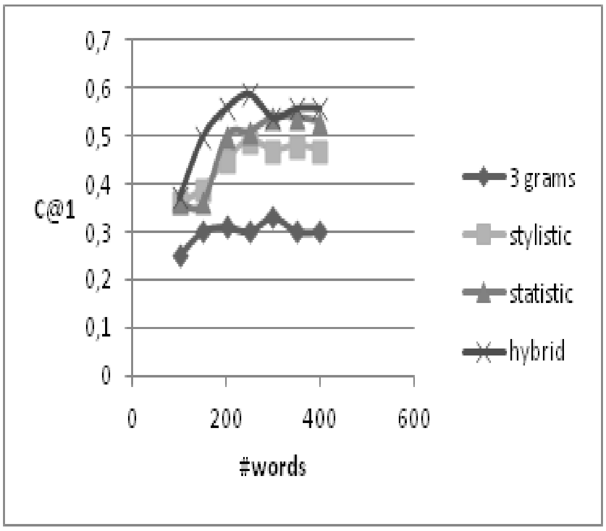


Figure 3: The C@1 Performance of different features according to words number

References

Stamatatos Efstathios, Daelemans Walter, Verhoeven Ben , Potthast Martin, Stein Benno, Juola Patrick, Miguel A. Sanchez-Perez, and Barrn-Cedeo Alberto. 2014. Overview of the Author Identification Task at CLEF. England.

Li Jiexun, Zheng Rong and Chen Hsinchun. 2006. From fingerprint to writeprint. Communication ACM 49(4), 7682.

Zheng Rong, Li Jiexun, Chen Hsinchun and Huang Zan. 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.

Vartapetiance Anna and Gillam Lee. 2014. A Trinity of Trials: Surreys 2014 Attempts at Author Verification. Proceedings of PAN@CLEF2014.

Argamon Shlomo, Whitelaw Casey, Chase Paul, Hota S. Raj, Garg Navendu and Levitan Shlomo. 2007. Stylistic text classication using functional lexical features Journal of American society of information science and technology 58(6), 802822.

Raghavan Sindhu, Kovashka Adriana and Mooney Raymond. 2010. Authorship attribution using probabilistic context free grammars. Proceedings of ACL10, 3842.

Feng V. Wei and Hirst Graeme. 2013. Authorship verification with entity coherence and other rich linguistic features.Proceedings of CLEF13.

Mccarthy M. Philip, Lewis A. Gwyneth, Dufty F. David and Mcnamara S. Danielle. 2006. Analyzing writing styles with coh-metrix. Proceedings of FLAIRS06, 764769.

Baayen R. Harald. 2008. Analyzing Linguistic Data. A Practical Introduction to Statistics using R.Cambridge, Cambridge University Press, Cambridge.

Mosteller Frederick and Wallace David. 1964. Inference in an Authorship Problem,1964. In Journal of the American Statistical Association, Volume 58, Issue 302, 275-309.

Labb Cyril. 2003. Intertextual Distance and Authorship Attribution. Corneille and Molire, In: Journal of Quantitative Linguistics, , pp. 213-231.

Burrows John. 2002. Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship, In Journal Lit Linguist Computing.

Blei M. David, and Jordan I. Michael. 2004. Variational methods for the Dirichlet process. In Proceedings of the twenty first international conference on Machine learning ACM.

Hershey R. John, Olsen A. Peder and Rennie J. Steven. 2007. Variational Kullback Leibler divergence for Hidden Markov models. IEEE Workshop on Automatic Speech Recognition and Under standing.

Grieve Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. Literary and linguistic computing, 22(3),.251-270.

Savoy Jacques. 2012. Etude comparative de stratgies de slection de prdicteurs pour lattribution dauteur, COnfrence en Recherche dInformation et Applications CORIA. 215-228, France.

Stamatatos Efstathios, Fakotakis Nikos and Kokkinakis George. 2000. Automatic text categorization in terms of genre and author, Computational Linguistics, Volume 26,.471-495.

Lee C. Min, Mani Inderjeet, Verhagen Marc, Wellner Ben, and Pustejovsky James. 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.

Zhao Ying, and Zobel Justin. 2007. Searching with style: Authorship attribution in classic literature, In Proceedings of the Thirtieth Australian Computer Science Conference ACM Press, 59-68,Australia.

Moreau Erwan, Jayapal Arun, and Vogel Carl. 2014. Author Verification: Exploring a Large setof Parameters using a Genetic Algorithm Notebook for PAN at CLEF 2014. England.

Peas Anselmo and Rodrigo lvaro. 2011. A Simple Measure to Assess Nonresponse. In Proceedings Of the 49th Annual Meeting of the Association for Computational Linguistics, Vol.1, 1415-1424.

Frery Jordan, Largeron Christine, and Juganaru-Mathieu Mihaela. 2014. UJM at CLEF in Author Identification. PAN@CLEF2014. England.