Authorship Verification, Average Similarity Analysis

Daniel Castro Castro
CERPAMID, Cuba
daniel.castro@cerpamid.co.cu

Yaritza Adame Arcia
DATYS, Cuba
yaritza.adame@datys.cu

María Pelaez Brioso
DATYS, Cuba
maria.pelaez@datys.cu

Rafael Muñoz Guillena
Universidad de Alicante, España
rafael@dlsi.ua.es

Abstract

Authorship analysis is an important task for different text applications, for example in the field of digital forensic text analysis. Hence, we propose an authorship analysis method that compares the average similarity of a text of unknown authorship with all the text of an author. Using this idea, a text that was not written by an author, would not exceed the average of similarity with known texts and only the text of unknown authorship would be considered as written by the author, if it exceeds the average of similarity obtained between texts written by him. The experiments were realized using the data provided in PAN 2014 competition for Spanish articles for the task of authorship verification. We realize experiments using different similarity functions and 17 linguistics features. We analyze the results obtained with each pair function-features against the baseline of the competition. Additionally, we introduce a text filtering phase that delete all the sample text of an author that are more similar to the samples of other author, with the idea to reduce confusion or non-representative text, and finally we analyze new experiments to compare the results with the data obtained without filtering.

Keywords: Authorship detection, Author identification, similarity measures, linguistic features.

1 Authorship Analysis

Determine the true author of a document has been a task of social interest from the moment it was possible to attribute the authorship of words. Questions about the authorship of a document may be of interest not only to specialists in the field (forensics specialist, linguistics researchers, etc.), but also in a much more convenient sense for politicians, journalists, lawyers. Recently, with the development of statistical techniques and because of the wide availability of accessible data from computers, the authorship analysis automatically has become a very practical option.

There are many practical examples where the authorship analysis becomes the key to solve them. Suppose a malicious mail is sent using an email account belonging to someone else, which subsequently are accused of this fact, who is the author of the mail? It may happen that a person dies and there is a note that makes it seem that the person committed suicide, it really was a suicide note or was used to cover up a murder? It may be a document, say a digital newspaper that is altered so it cannot be used as evidence in a trial, was it or not altered this newspaper?

The authorship analysis task confronts the problem of determining the author of an anonymous document or one whose author is in doubt. For this it is necessary to try to infer linguistic characteristics (features) of the author through documents written by him, features that will allow us to create a model of the writing style of this author and measure how similar may be any unknown document to documents written by that author.

One of the principal evaluation labs for the dissemination, experimentation and collaboration in the development of methods for the authorship analysis is found in the PAN[1] lab associated to CLEF. It is important to notice, that most of the papers presented in different editions of this evaluation forum (Joula and Stamatatos, 2013; Stamatatos et al., 2014) used Natural Language Processing tools, in order to obtain the linguistic features which identify an author and differentiate it from the rest.

[1] http://pan.webis.de

Proceedings of Recent Advances in Natural Language Processing, pages 84–90,
Hissar, Bulgaria, Sep 7–9 2015.

In PAN editions, 2013 and 2014, specifically it was tested the task of authorship verification, where authors samples are formed by known author documents and an unknown document to check whether it was written by that author. No restrictions is imposed on the use of samples of others for support in finding a decision, or just use the samples of single author, the latter idea would be challenging and difficult because we need to capture the writing style of the author only with his samples.

The basic properties of the papers presented in the PAN 2014 authorship verification task (Stamatatos et al., 2014) are:

1. By the use of known documents samples of authors: intrinsic (only the documents of the author in analysis) or extrinsic (using samples of others authors).
2. Type of machine learning algorithms or approximation used: lazy or hard-working approaches (more training computational costs).
3. Type of linguistic features used: low-level features (characters, phonetic and lexical) and/or syntactic.

1.1 Linguistic Features

The *linguistic features* are the core of the authorship analysis task (regardless of the subtask or approach used in the analysis, such as author verification, author detection, plagiarism detection, etc.), they can be used to coded documents with any mathematical model, traditionally being the vector space model the approximation most used. The purpose lies in trying to identify a writing style of each author to distinguish it from the rest (Juola, 2008).

There are several number of features that have been taken into account in the authorship analysis task, in the majority is used a distribution of features grouped by linguistic layers (we call them also features obtained from the content writing) (Ruseti and Rebedea, 2012; Halvani et all., 2013; Castillo et all., 2014; Khonji and Iraqi, 2014).

Five linguistic feature layers are identified in (Stamatatos, 2009): phonetic, character, lexical, syntactic and semantic layer:

1. Phonetic layer: This layer includes features based on phonemes and can be extracted from the documents through dictionaries. Example: the International Phonetic Alphabet (IPA).
2. Character layer: This layer includes character-based features as prefixes, suffixes or n-grams of letters.
3. Lexical layer: This layer includes features based on terms such as auxiliary words.
4. Syntactic layer: This layer includes syntax based features such as sentences components.
5. Semantic layer: This layer includes semantic-based features as homonyms or synonyms.

Based on this structure feature layers, in our present work we use features of the 2,3 and 4 layers, which we illustrate in more detail in next sections.

In Section 2 we present the characteristics of our method and in section 3 the experimental results using the data of Authorship Verification PAN 2014 competition. Finally conclusions and future work.

2 Average Similarity Proposal

There are various aspects that need to be analyzed in order to implement a method that allows us to assess whether a text of unknown or disputed authorship, was written by an author from which we have written sample texts. It should be considered whether samples of the author belong to the same genre, theme, were written with a considerable time difference, are written in the same language or have sections written in other languages, or if the samples have been revised and corrected by someone else.

From a practical point of view in software application (real scenario) for the algorithms we also do not have the assurance that all documents given as examples of an author, have actually been written by the author in question. That is, it is possible that some samples were drafted by someone else.

Our method is based on the analysis of the average similarity (AS_{Unk}) of an unknown authorship text with the closeness to each of the samples of an author, comparing it to the Average Group Similarity (AGS) between samples of an author.

We performed experiments with a total of 17 types of linguistic features (we will illustrate the features in the following section) and used six similarity functions.

We identified three key steps in our method, these are:

1. Representation of all documents by one feature type. This must be done for all the features.

2. Average similarity between the documents samples of an author (AGS).
3. Average similarity between the document of unknown authorship and the known samples of one author (AS_{Unk}).

2.1 Linguistic Features used to Represent the Documents

We use the vector representation to store the values of the linguistic features extracted form one document, so each sample (document) with known or unknown author is represented by 17 vectors corresponding to each of the types of features with which experiments were performed.

The features evaluated and calculated are grouped in three layers: character, word and syntactic (lemma and Part of Speech)

1. Character
    a. Tri-grams of characters
    b. Quad-grams of characters
    c. Uni-grams of prefixes of size 2
    d. Uni-grams of suffixes of size 2
    e. Bi-grams of prefixes of size 2
    f. Bi-grams of suffixes of size 2
2. Words
    a. Uni-grams of words
    b. Tri-grams of words
    c. Bi-grams of words at the beginning of sentence
    d. Punctuations marks
3. Lemma and Part of Speech
    a. Uni-grams of lemmas
    b. Uni-grams of Part of Speech
    c. Tri-grams of lemmas
    d. Tri-grams of Part of Speech
    e. Bi-grams of lemmas at the beginning of sentence
    f. Bi-grams of Part of Speech at the beginning of sentence

The features of the third layer of analysis are obtained using tools of Natural Language Processing implemented in the *Xinetica*[2] platform.

2.2 Average Similarity

To illustrate the performance of our method, we show in the Figure 1 the process to calculate the average similarity from the documents of the known author and the average similarity of these samples with the unknown text. Initially we have several samples of documents (Doc) by an author and a document of unknown authorship (Unk).

The first task is to represent each of these documents in a vector space model, analyzing one type of feature. Subsequently, for the samples documents of the author we analyze the average similarity of each document with the rest, using the following formula:

$$AS_j = \frac{\sum_{o_j \in K_j} Sim(O, O_j)}{|K_j| - 1}$$

Where "O" would be a document of the author and "O_j" the rest of the documents of the same author, K_j represents the author and $|K_j|$ the number of documents of the author. By $Sim(O, O_j)$, it´s represented the similarity between two documents.

Therefore, for each document of known author their average similarity with the other is calculated and finally, the average similarity of all samples is calculated or what we call the average group similarity (AGS):

$$AGS = \frac{\sum_{o_j \in K_j} AS_j}{|K_j|}$$

Given document of unknown authorship, initially must be represented by the type of feature in which samples of known author are represented with which are to be compared. Then the AS_{Unk} is calculated using the known samples. The decision is made by comparing the AGS with unknown calculated AS_{Unk}. If $AS_{Unk} < AGS$, then the unknown sample is not considered written by this author. To determine if the response is positive (that is, that the document of unknown author was written by the author of the given samples), then the $AS_{Unk} \geq AGS$.

We have implemented 6 similarity functions in order to perform experiments with each of them, these are: Cosine, Dice, Jaccard, Tanimoto, Euclidean and MinMax (Gomaa and Fahmy, 2013).

One element to prove that we incorporate is related to the analysis of samples of each author, in order to filter out those that do not represent or characterize the writing style of the author. We incorporated a filtering stage prior to the calculation of AGS.

For each sample, the AS was calculated for each group of samples of the authors and eliminates those samples of documents that had an AS value greater with samples of different authors to his corresponding author. This filtering variant we will call "Non typical" and the variant without

[2]http://www.cerpamid.co.cu/xinetica/index.htm

filtering its call "No reduction". This reduction variant for not typical documents would be good in the future to test the effect or impact it would have on different collections of texts of the authors. For example, how it would affect the analysis of authorship if the authors samples correspond to the same topic or even an author's samples were not of the same length or a single topic.

We focus then our study in analyzing three aspects:

1. The idea of the AGS measure as a limit to determine when an unknown document was written by an author. We see this as an intrinsic approximation to the task.
2. Non typical known documents eliminated don't affect the purpose of correct identification the author of an unknown one, or incorrect assigning to an author a text that was not written by him.
3. How far are the results of each pair function-features in correspondence with the best and baseline of the experiments reported in PAN 2014 competition for Spanish dataset, in order to evaluate if the AGS measure could be used.

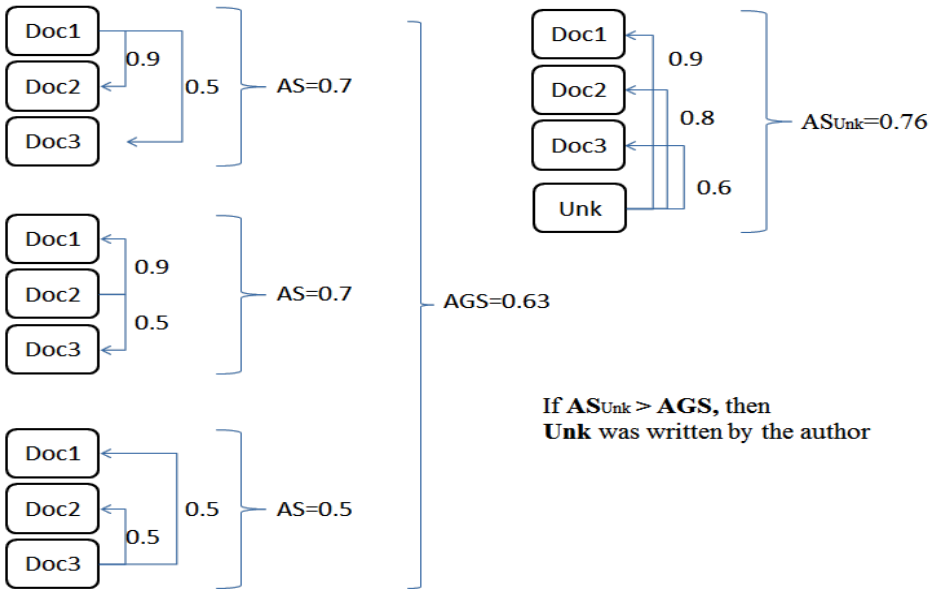

Figure 1: Average Group Similarity (AGS) analysis of an author documents samples and Average Similarity (AS_{Unk}) of an unknown authorship document

3 Experimental Results

With each pair function-feature we would evaluate the authorship verification method we propose. This section shows the results of evaluating the training and test dataset offered in the task of authorship verification of the PAN 2014 edition for the Spanish language using the *accuracy* measure. We present the results for each pair function-feature without reducing known documents samples of the authors and using a filtering phase where *Non typical* documents are eliminated.

In train and test dataset there are a maximum of 5 documents samples for each author and one unknown text, and the purpose is to determine if this unknown sample was written by this author. The train data has 100 authors and the test data 50.

The evaluation measure we use is *accuracy c@1* (Peñas and Rodrigo, 2011). This is the measure used in the competition:

$$c@1 = (1/n)*(nc+(nu*nc/n))$$

where n is the number of problems that correspond to the number of authors, nc is the number of correct answers (i.e. say **not** written by the author when the unknown text was indeed not written by him and **yes** when it was written)

and *nu* is the number of unanswered problems. In our method we answer all the problems so the *nu* value would be 0 and then we would evaluate *accuracy = nc/n.*

In (Stamatatos et al., 2014) are presented all the details of the dataset for the languages evaluated in the competition. In the overview is presented a baseline *accuracy* value that allows us to evaluate and compare the results of the participants, the *accuracy* value is 0.53 for the Spanish data. The best value of accuracy obtained in the competition was 0.79 using a META-CLASSIFIER developed with the combination of all the results of the participants. The best accuracy of a participant method was of 0.77 achieved by (Khonji and Iraqi, 2014).

Figures 2, 3 and 4 show the results with the test data with and without reduction, that is, in Figure 2 the results for all features of the character layer of are shown with and without reduction and likewise for 3 and 4. For most pair function-feature and both variants reducing samples or not, the values obtained with the test data are greater than the values obtained with the training, but its observed a uniform behavior with respect to those achieved with the test data.

As a general rule, with the features of the Character layer, the best results are appreciated for representations based on n-grams of characters for n 3 and 4; as well as the bi-grams of prefixes and suffixes of words. With regard to the similarity functions, highlight the values obtained using Dice and Jaccard, being quite similar.

If we analyze the results according to filtering variant of the samples, it is observed that the values of accuracy are slightly higher with the analysis of *Non typical*, the difference would lie in the need for a greater effort in the previous stage in which the non-typical samples are filtered, but for classification of unknown texts it would need less computing time.

In Figure 3 are appreciated the results without reducing samples and non-typical samples reduction for features of the layer Words.

We evaluate as positive the values achieved with representations of n-grams of words with *n* 1 and 3, noting that for uni-grams of words with the functions Dice and Jaccard are achieved the best values (0.78 and 0.8 of accuracy in that order) in all tests with any of the features from the three layers and close to the best obtained in the PAN 2014 competition for the Spanish dataset which was accuracy 0.79 from a meta-classifier (Stamatatos et al., 2014).

The Euclidean and MinMax functions (dissimilarity functions), in most cases have the lowest values.

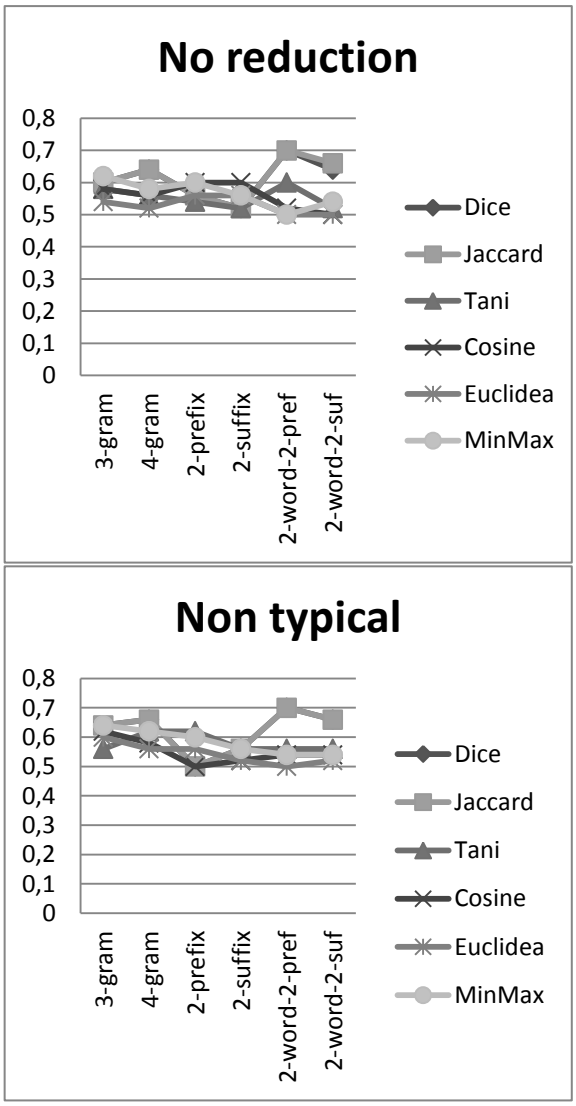

Figure 2: Results for the Character layer of features and all the similarity functions. No reduction filtering and Non typical reduction.

Figure 4 shows the results for the features of the Lemma and Part of Speech (PoS) layer.

There are illustrated good values with the representations of lemmas and PoS n-grams for *n* 1 and 3, primarily working with lemmas. It can be noted that to each word correspond a lemma and for one lemma may be associated more than one word, taking this into account we can analyze the results using the lemma n-grams and word n-gram representations.

For example, we see that for the variant without reduction of the samples, the results with the representation of words (terms) are higher

compared to the use of lemmas, and very similar if we use the feature representations 3-grams of words or lemmas. For variant with non-typical reduced samples, the results were quite similar for any of the representations.

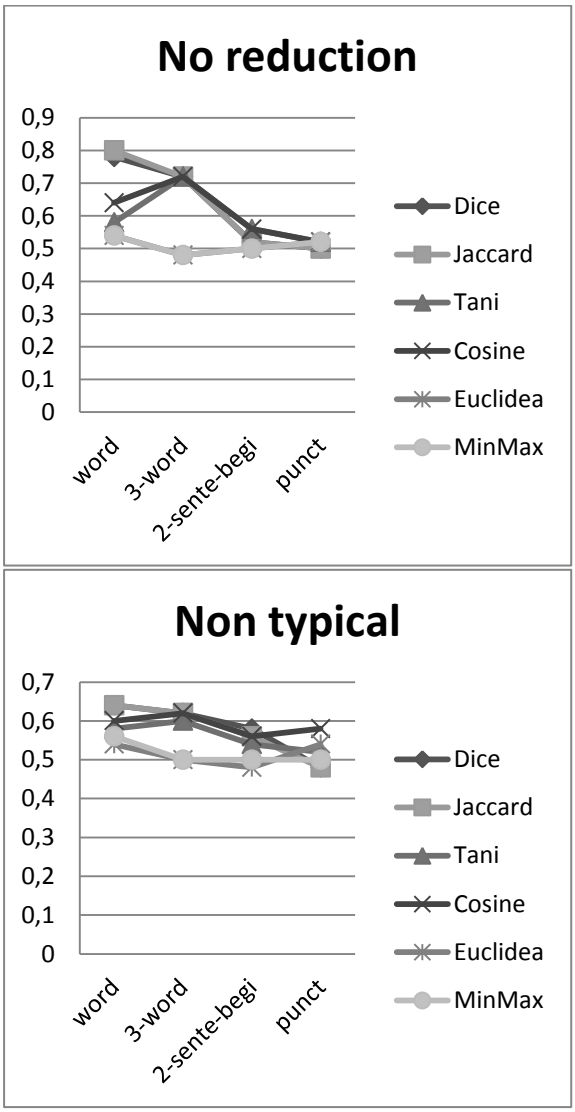

Figure 3: Results for the Word layer of features and all the similarity functions. No reduction filtering and Non typical reduction.

To summarize, the best results over the baseline value is obtained using the functions Dice, Jaccard, Tanimoto and Cosine, from these Dice and Jaccard are highlighted.
Analyzing the features representations used, good values are obtained with several features and especially those in which are achieved accuracy values close to 0.7 or higher.
Regarding to the reduction variants of samples texts of the author's, with some pair's function-features, are obtained better results without reducing samples and in other cases by non-typical filtering.

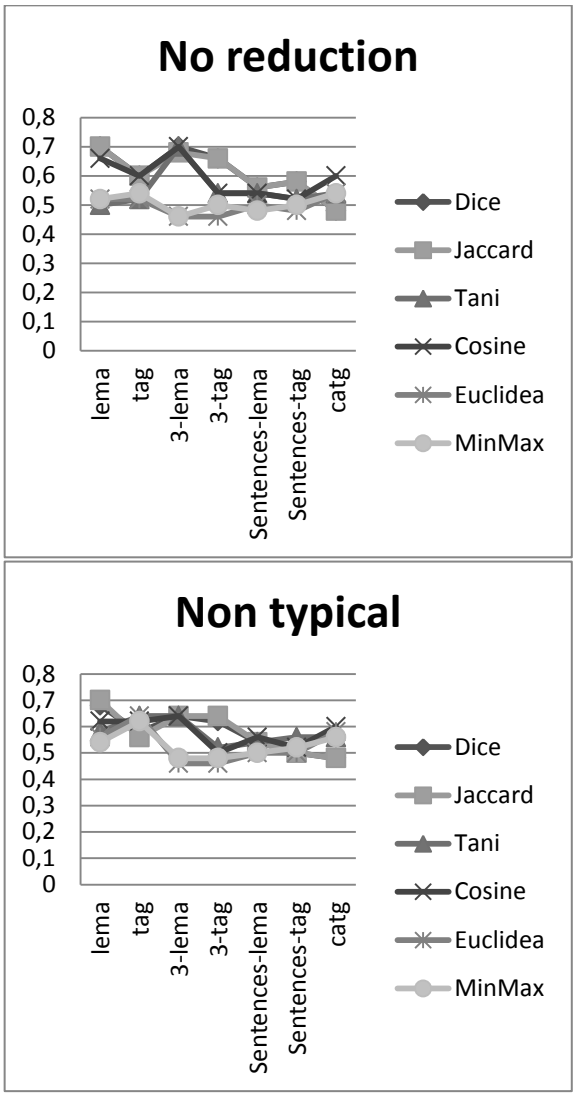

Figure 4: Results for the Lemma and Part of Speech layer of features and all the similarity functions. No reduction filtering and Non typical reduction.

4 Conclusions and Future Work

We have presented the implementation of a method for authorship analysis that compares the average similarity calculated between a document of unknown authorship and documents written by an author, with the average similarity of the samples of this author.
Using this idea, a text that was not written by an author, would not exceed the average of similarity with known texts and only the text of unknown authorship would be considered as

written by the author, if it exceeds the average of similarity obtained between texts written by him.
To prove the idea, we use 17 types of linguistic features to represent the documents and evaluate the similarity between two vector representations of documents using one of six's similarity functions implemented. We tested the method with each pair function-feature, evaluating the results between each execution and taking into account the baseline and best results exposed in the authorship verification task with training and test data of the PAN 2014 for the Spanish edition.
We also include a preliminary phase for reducing samples texts of each author, with the intention that the samples of the authors were representative of his style of writing and little similar to the samples of other authors, calling these *Non typical* reduction.
We evaluate the results of each pair function-feature without reducing samples and for *Non typical* reducing. This allowed us to assess whether occurred a drastic reduction in test results when samples of texts written by an author are eliminated, ensuring that the results do not differ much and in some cases increase.
We obtained several results above the baseline value reported in competition and in some cases near to the best.
We propose as future work, the implementation of a method that allows us to combine several function-feature pair's in order to give a final conclusion with some voting mechanism.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Science and Innovation (TIN2012-38536-C03-03)

References

Castillo, E. Vilariño, D. Pinto, D. León, S. Cervantes, O. *Unsupervised method for the authorship identification task*. Notebook for PAN at CLEF 2014.

Gomaa, W. and A. Fahmy. 2013. *A Survey of Text Similarity Approaches*. International Journal of Computer Applications (0975 – 8887) Volume 68– No.13.

Halvani, O. Steinebach, M and Zimmermann, R. *Authorship Verification via k-Nearest Neighbor Estimation*. Notebook for PAN at CLEF 2013.

Juola, P. 2008. *Authorship Attribution*. Foundations and Trends in Information Retrieval Vol. 1, No. 3 (2006) 233–334

Juola, P. and E. Stamatatos. 2013. *Overview of the Author Identification Task at PAN 2013*. CLEF 2013.

Khonji, M and Iraqi, Y. *A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF)*. Notebook for PAN at CLEF 2014.

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, pages 1415-1424.

Ruseti, S and Rebedea, T. 2012. *Authorship Identification Using a Reduced Set of Linguistic Features*. Notebook for PAN at CLEF 2012.

Stamatatos, E., W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola and M. Sanchez-Perez. 2014. *Overview of the Author Identification Task at PAN 2014*. CLEF 2014.

Stamatatos, E. 2009. *A Survey of Modern Authorship Attribution Methods*. Journal of the American Society for Information Science and Technology, 60(3), pp. 538-556, 2009, Wiley.