An Approach to Detecting Text Autorship in the Spanish Language

Mauricio Iturralde, Roberto Maldonado and Daniel Fellig

Universidad San Francisco de Quito (USFQ)
miturralde@usfq.edu.ec, dfellig@usfq.edu.ec,
roberto.maldonado.galiano@estud.usfq.edu.ec

***Abstract*—Authors tend to express themselves using language in ways that reflect particular styles, vocabularies, biases, idioms, etc. These features can be captured in the so-called firm or stylone. Although capturing these attributes with high fidelity has proven to be very challenging, some advances have been made. Stylometry is the analysis of the unique attributes that are expressed by an author unconsciously through his or her publications.**

In this paper we investigate techniques for the detection of authorship patterns from the text content of a large number of digital documents, including e-mails, academic notes and free redaction in the Spanish language. A mechanism based on pondering parameters, including statistical observations, extracting a pattern is proposed. We defined 150 stylistic criteria parameters adapted to the Spanish language to compute our metric.

Extensive experiment results are also presented.

***Keywords*—Stylometry, writing, authorship, semantic, syntax**

I. Introduction

In the last years, social networks have grown massively and become a powerful tool for the masses to broadcast messages. Activist forces of different types such as social, political, economic, or environmental use social networks to send their messages to the masses. Obviously, social networks can also broadcast messages from anonymous sources that at times must be investigated for authorship by governments, intelligence agencies and homeland security departments for security reasons. An obvious example could be when a country's leader's life is threatened.

Authorship detection or attribution is the process whereby the most likely author of an anonymous document is identified based on a collection of known documents. In this context, a text whose author has not yet been identified is compared in some way with other texts of known authors to identify authorship, or at least to exclude some list of authors.

According to [1], stylometry is based on the following major premises: (a) authors express a unique, consistent, recognizable style in their writing, (b) writing style can be quantified and extracted from written works, (c) writing style can be statistically or otherwise computationally analyzed, and (d) this analysis allows for profiling authors such that they can be recognized automatically. The recognition (or verification) of an author via their style is authorship attribution or identification. There is a great variety of characteristics used to distinguish authors, including word frequencies, formatting choices, spelling errors, and punctuation.

However, due to cultural differences, all those characteristics might not be equally applicable for every language.

A system able to attribute a certain text to a particular author has to be trained with some texts from that author. From those texts, the system has to extract metrics to identify patterns that are present in other texts from the same author, but not in texts whose authorship is different. In this sense, it is a particular case of one-class learning in which several positive examples are presented to the system, but no instance of a non-member example is presented.

In this paper we present a first approach of stylometry for authorship recognition in the Spanish language. We have selected several parameters that are used in stylometry in the English language, and have added new parameters to identify writing style in the Spanish language to use in our authorship detection method.

The remaining parts of this paper are presented as follows: In Section II we present a state-of-the-art of Stylometry. In Section III we present the parameters that we selected to compute our metric. In Section IV we introduce our proposed method and compare it against the Minkowski and Chebishev methods. In Section V we show our numerical results, and finally in Section VI we present our conclusions and future work.

II. Literature Review

High accuracy is demanded from authorship identification methods when being used as digital evidence in courts of law for criminal offenses and civil disputes. Authorship analysis techniques have been used to identify the gender, age, ethnic origin, and education level of authors, even when this information was not provided by the authors. This process focuses on profiling authors, and then being able to discriminate between them to decide on the authorship of a written document. Authorship analysis can be applied to many variants of this problem including, authorship identification, authorship verification and authorship characterization problems [3].

There are several methods that have been proposed to accomplish this task for the English language.

In [2], authors claimed that their proposed method has obtained 95% accuracy and was successfully used in several digital crimes. These methods are applied to documents that

978-1-5090-2914-3/16/$31.00 ©2016 IEEE

are strongly thematic, involving letters of apology, or writing expressing a terrible situation experienced by the author. These writings are then analyzed with lexical and semantic methods to produce various statistical results. These results can then be complied and compared with the disputed text to obtain the final result.

In [4], the authors used a type of authorship analysis framework which involves the extraction of three types of message features namely; style markers, structural features and content-specific features. Another method which according to the authors has achieved very high success rates is presented in [5]. This method focuses on the use of a Bayesian classifier with stylometric features and function words. They have explored various methods with lexical and syntactic feature sets, with the Bayesian classifier with Gaussian density yielding the best results after applying Principal Component Analysis (PCA). They tested their methods using published documents on a news website from 18 different authors, each had written more than 500 articles. Lexical and stylometric features such as number of words, average word length, vocabulary size and number of punctuations are extracted. Their findings show that high success rates can be achieved when there are a lot of data available for analysis. Hence, the success rate will increase with the number of available samples.

In [6], the authors propose a technique for authorship similarity detection using text content of a short, subject-free email, extracting a common pattern and applying machine learning. 150 stylistic cues are identified for this problem. According to the authors, this method can achieve up to 89% of accuracy when each identity has between 10 and 15 short emails respectively.

In [7], the authors introduce a mechanism that works as an hybrid model based on self-organizing maps and in information-theoretic aspects. In the proposed solution, a mutual information function of unknown texts are compared to the mutual information function of texts from a known author. If the distance between these two distributions exceeds a certain threshold, then the unknown text is from a different author, otherwise the authorship is the same.

Another method is presented in [8]. This method is based on representations of users and documents for grouping and authorship identification. It works in a two-layer framework that enables to apply authorship identification over larger number of authors (100). The proposed two-layer solution divides the large number of authors into smaller groups that contain reasonable numbers of authors (5 - 14 authors) and modify the classification to performed across two stages by attributing first the appropriate group and then identifying the particular author within the group.

In [9], an algorithm for building best feature sets based on a test corpus, as well as some considerations for feature sets in problem spaces for author attribution in more than one language or character set is presented. This study was focused on the Russian and English languages.

In [10], authors presented a methodology based on a multi-objective genetic algorithm and Support Vector Machine classifier to select the most discriminating subset of syntactic attributes for authorship attribution. Experiments were made on a database composed of 3000 short articles written in Portuguese.

III. Parameters Selection

The proposed mechanisms exposed in Section II are implemented based on a list of parameters. In Section 2 all of the proposed mechanisms were created for the English language. None of them tackled this topic for the Spanish language. As every language has its its own rules, the parameters must be chosen according to those rules. Our research focuses on the Spanish language; therefore, here we present the list of parameters used by our method.

In Table I we classify the parameters into groups: Punctuation, mathematical, and symbols.

Table I: General Characters

Punctuation Group	
PPD:	Punctuation Period
PCN:	Punctuation Colon
PCA:	Punctuation Comma
PSC:	Punctuation Semicolon
EEO:	Expression Exclamation Open
EEC:	Expression Exclamation Close
EQO:	Expression Question Open
EQC:	Expression Question Close
Mathematical Group	
MPS:	Mathematical Plus
MMS:	Mathematical Minus
MTS:	Mathematical Times
MMO:	Mathematical Modulus
MEQ:	Mathematical Equals
MGR:	Mathematical Greater
MLR:	Mathematical Lesser
MDN:	Mathematical Division
Group Symbols	
GPO:	Grouping Parenthesis Open
GPC:	Grouping Parenthesis Close
GCO:	Grouping Curly Bracket Open
GCC:	Grouping XCurly Bracket Close
GBO:	Grouping Bracket Open
GBC:	Grouping Bracket Close
Other Symbols	
OBS:	Other Back Slash
OAA:	Other Arroba
ONS:	Other Number Symbol
ODS:	Other Dollar Sign
OHO:	Other Carret
ONP:	Other Ampersand
OVD:	Other Pipe
OUS:	Other Underscore

In Table II we list the accented vowels which are an important part of the Spanish language structure. In Table III we list the numbers, and in Table IV we list the most common Spanish greetings and farewells. We consider the Spanish greetings as an important parameter set to be considered in our method.

The parameters that were selected can be classified into various groups depending on what is intended to establish about the author:

- Lexicon analysis; this analysis seeks to determine the manner which an author uses a certain set of words within a writing segment.
- Analysis of syntax tendencies; the focus is on the permutations in the usage of sentences within paragraphs,

Table II: Accents

MAA:	Accented Upper Case A
MAE:	Accented Upper Case E
MAI:	Accented Upper Case I
MAO:	Accented Upper Case O
MAU:	Accented Upper Case U
MIA:	Accented Lower Case A
MIE:	Accented Lower Case E
MII:	Accented Lower Case I
MIO:	Accented Lower Case O
MIU:	Accented Lower Case U

Table III: Numerical

CERO:	0
UN, UNO:	1
DOS:	2
TRES:	3
CUATRO:	4
CINCO:	5
SEIS :	6
SIETE:	7
OCHO:	8
NUEVE:	9

Table IV: Spanish Greetings And Farewells

G1:	Hola
G2:	Buen Dia
G3:	Estimado
G4:	Estimada
G5:	Buenos
G6:	Buenas
F1:	Gracias
F2:	Atentamente
F3:	Cordialmente
F4:	Suerte
F5:	Saludos

words in sentences, and letters in words, including the distribution and frequency of punctuation marks.

- Distribution of letters analysis; the extraction of the number of each letter used. This analysis determines the frequency of usages of upper case letters in relation to usages of lower case letters when texts are compared. Included are also the usages of lower and upper case vocals that are accented as defined in the usages of the tilde.
- Usage of numbers expressed as words versus expressed as numerical digits; this behavior determines another stylistic quality in the writing that can be used to differentiate authorship.
- Detection of degree of presence or absence of words that express greetings and farewells; this can depict the degree of formality as well as the general attitude the author has towards the person or institution being addressed. It must be taken in context with the type of social environment the author operates in his or her daily life. This will surely determine the language in general that the author allows him or herself to use.

IV. Proposed Solution

In this research, machine learning methods are not used to generate a predictive model or classifier based on sets of the selected parameters. This exclusion is made on the basis that Bayesian classifiers are supported on the often erroneous assumption that each of the attributes are equally important and mutually independent [11]. In the case of stylometry, data clearly shows the dependence of stylometric attributes amongst themselves clearly invalidating the Bayesian classifier assumption. In this research we present an approximation based on the weighing of the arithmetic mean of parameters based on their importance. A heavier weight is granted to the parameters that appear more frequently among all texts.

Let us consider $T = \{t_1, t_2, t_3, \ldots, t_n\}$ as the set of texts t of a known author, this set of texts are used to obtain a writing pattern that will be compared against another text of an unknown author that we call t_u.

Let us also consider $P = \{p_1, p_2, p_3, \ldots, p_m\}$ as the set of parameters that are used to obtain the writing pattern of the known author's texts and the unknown author's text; e.g., the correct use of the symbol '?' or the percentage of used Capital Letters.

By using Formula 1, the average of p_i is computed based on every text of T

$$\overline{p_i} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{1}$$

As result we have the set $\overline{P} = \{\overline{p_1}, \overline{p_2}, \overline{p_3} \ldots \overline{p_m}\}$

The set $P' = P$ are the parameters of the unknown authorship text t_u. Thus the distance between the sets $\overline{P}$ and P' is computed by using Formula 2 giving as result the set $D = \{d_1, d_2, d_3 \ldots d_m\}$

$$d_i = \frac{|\overline{p_i} - p'_i|}{|\overline{P}|} \tag{2}$$

The process of matching the estimation between P and $\overline{P}$ will be computed by using the values of the set D. Those values will be used in the four methods presented in the next section.

A. Simple proportion method

The first approach used for computing the matching between the known and unknown texts is basically a proportion method using Formula 3.

$$\delta = \sum_{i=1}^{m} \frac{(|1 - d_i| \times 100)}{|D|} \tag{3}$$

The value of δ shows the matching percentage of texts.

B. Pondered proportion method

This method modifies Equation 3 by adding a weight $W = \{w_1, w_2, w_3, \ldots, w_m\}$ where $\sum_{i=1}^{m} w_i = 1$ and $w_i \geq 0$.

The matching percentage value Υ of this method is computed by using Formula 4.

$$\Upsilon = \sum_{i=1}^{m} \frac{(|1 - d_i| \times 100)}{|D|} \times w_i \tag{4}$$

where

$$w_i = \frac{p_i}{\sum_{i=1}^{n} p_i} \tag{5}$$

The value p_i is obtained from the frequency rate of every parameter.

This approach seems to be more reasonable due to the fact that all parameters cannot be considered of equal importance. All the investigated parameters for finding a pattern of authorship are not homogeneous, therefore it is important to assign specific weights to each parameter.

C. The Minkowski Distance

This is a well-known mathematical tool used to quantify the level of how well two vectors match [12]. In our approach we define two vectors corresponding to the average of parameters $\overline{P}$ associated with the known author, and the average of the parameters P' associated with the unknown author.

The matching percentage is computed by using the distance concept. The Minkowski distance of p order for two vectors is represented by Formula 6.

$$\Gamma = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{\frac{1}{p}} \tag{6}$$

By the definition of Minkowsky inequality, the value p must be ≥ 1

D. The Chevishev Distance

The Chevishev distance Ω (Formula 7) is a variant form of Equation 6.

$$\Omega = \lim_{p \to \infty} \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{\frac{1}{p}} \tag{7}$$

The variant occurs when the value $p \to \infty$. The values lower than the max value are not taken into account, so the Equation 7 becomes:

$$\Omega_{(A,B)} = max(d_i) \tag{8}$$

$$d_i = |a_i - bi| \tag{9}$$

Therefore, the metric represents basically the higher distance between two points excluding lower distances from different points [13].

V. Numerical Results

Our proposed method has been tested by using a large number of digital documents, including e-mails, academic notes and free redaction. All tested documents were obtained from 35 different authors. The authors had no idea that their writing would be considered for pattern analysis, therefore we assume that the analyzed documents were written in styles natural to the authors with no artificial preconceptions.

Table V: Test environment

Authors	350
Known author's documents group	5
Unknown author's documents	1
Number of tests	350

In every test, one document by a known or unknown author was compared against five documents by a known author. In our test environment, unknown author can be considered as the same five text author or a different one.

We performed our test under this rule way because we consider that in order to evaluate our method, both results (the same and different author) are needed. So, let us consider R_s as the result of the test when the unknown text belongs to the same author and R_d as the result of the test when the unknown text belongs to a different author.

The main factor that help us to evaluate an authorship match is the value Φ that is the distance between R_s and R_t that we call our decision interval as shown in Formula 10.

$$\Phi = R_s - R_d \tag{10}$$

In order to have a good support for making our final decision when judging about a text authorship Φ must be as high as possible.

Table VI: Average of Results

Applied Method	Decision Interval Φ
Simple proportion	10 %
Pondered Parameters	41 %
Minkowski Distance	24 %
Chevishev Distance	5 %

As seen in Table VI, the average of the performed test show that the Pondered Parameters is the method that shows a highest performance. Making a decision with more than 40 percentage points seems to be a good deal for this kind of problems.

The second method that shows good results is the Mikowski Distance with a decision-interval average of 24 %. The simple proportion and the Chebishev Distance cannot be considered as solution for this problem due to those bad results. Figure 1 show the average performance obtained when testing all four methods.

The pondered parameters method performs the best results due to the fact that it does not consider all the parameters as equals. When authors write their text, they do not always reflect all the parameters, sometimes they barely reflect any

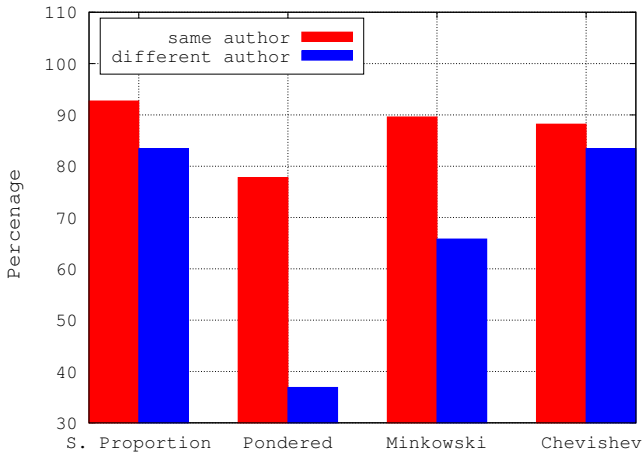

Figure 1: Performance of similarity of all methods

parameters. Rarely used parameters tend to affect the average result. We consider that is not wise to compare some texts where some parameters appear often against others where parameters appear barely. This is the main argument for pondering parameters when making the final decision.

When all the texts from a same author or not are compared, our method discriminates the non common parameters by setting their weights to 0.

VI. Conclusions and Future Work

In this paper we introduced a mechanism to identify text authorship in the Spanish language. Our method proposes using several parameters with weights based on their periodicity. We quantified and tested our mechanism using the well known Minkowski and Chevishev distances.

We can conclude that an optimal method must not only obtain a high score when text of a given author is compared to text of the same author, but also obtain a low score when text does not belong to the same author.

The results of our tests show that our proposed method gives the best results. This good performance can be attributed to the fact that in our method we considered all parameters as heterogeneous.

Future work can focus on using our mechanism with another set of pondering parameters. Also, the determination of authorship of short text messages in social networks could be another interesting field for future work.

References

[1] L. M. Stuart, S. Tazhibayeva, A. Wagoner, and J. M. Taylor. "On Identifying Authors with Style ". *in 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2013. Manchester, UK,

[2] C. E. Chaski "Who's at the keyboard? authorship attribution in digital evidence investigations ". *in International Journal of Digital Evidence*, 2005.

[3] E. Stamatatos "Authorship attribution based on feature set subspacing ensembles ". *International Journal on Artificial Intelligence Tools*, pp. 1–16, 2006

[4] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang "A framework for authorship identification of online messages: Writing-style features and classification techniques ". *Journal of the American Society for Information Science and Technology*, vol. 57, pp. 378–393, 2005

[5] I. Bozkurt, O. Baghoglu, and E. Uyar "Authorship attribution ". *International Symposium in Computer and Information Sciences (ISCIS)*, pp. 1 - 5, 2007

[6] Xiaoling Chen, Peng Hao, R. Chandramouli, K. P. Subbalakshmi "Authorship Similarity Detection from Email Messages ". *in Springer Lecture Notes in Computer Science* , vol. 6871, pp. 375-386

[7] Antonio Neme, Blanca Lugo, Alejandra Cervera "Detection of Different Authorship of Text Sequences through Self-organizing Maps and Mutual Information Function". *in Springer Lecture Notes in Computer Science* , vol. 6838, pp. 186-195

[8] Haytham Mohtasseb and Amr Ahmed "Two-layer classification and distinguished representations of users and documents for grouping and authorship identification". *in IEEE Int. Conf. Intelligent on Computing and Intelligent Systems ICIS*, vol. 1, pp. 651-657, 2009 Shanghai, China

[9] Lauren M. Stuart, Saltanat Tazhibayeva, Amy R. Wagoner, and Julia M. Taylor "Style Features for Authors in Two Languages". *in IEEE/WIC/ACM Int. Conf. on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* , vol. 1, pp. 459 - 464, 2013 Atlanta, USA

[10] Paulo Varela, Edson Justino, and Luiz S. Oliveira "Selecting Syntactic Attributes for Authorship Attribution". *in Proc. of Int. Joint Conference on Neural Networks* , pp. 167 - 172, 2011 San Jose, CA, USA

[11] Brett Lantz "Machine Learning with R: Learn how to use R to apply powerful machine learning methods and gain an insight into real world applications". *in Livery Place: Packt Publishing Ltd.* , Packt Publishing Ltd

[12] R. Kamimura, Kanagawa and O. Uchida "Greedy network-growing by Minkowski distance functions ". *Proc. IEEE International Joint Conference on Neural Networks*, vol 4. pp. 2837 - 2842, Jul 2004 Budapest, Hungary

[13] Torleiv Kløve, Te-Tsung Lin, Shi-Chun Tsai and Wen-Guey Tzeng. "Permutation Arrays Under the Chevishev Distance ". *IEEE Transactions on Information Theory*, vol 56(6). pp. 2611 - 2617, Jun 2010 Budapest, Hungary

[14] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, and John Bethencourt "On the Feasibility of Internet-Scale Author Identification ". *IEEE Symposium on Security and Privacy*, pp. 300 - 314, May 2012 San Francisco, USA

[15] Maxim Shevertalov, Jay Kothari, Edward Stehle, and Spiros Mancoridis "On the Use of Discretized Source Code Metrics for Author Identification ". *IEEE International Symposium on Search Based Software Engineering*, pp. 69 - 78, May 2009

[16] Paulo Varela; Edson Justino; Luiz S. Oliveira "Selecting syntactic attributes for authorship attribution ". *IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 167 - 172, May 2011