An evaluation measurement in automatic text classification for authorship attribution

Una medida de evaluación en la clasificación automática de texto para atribución de autoría

Medida de avaliação na classificação automática de texto para atribuição de autoria

Para citar este artículo / To reference this article / Para citar este artigo: Rico-Sulayes, A. (2015). An evaluation measurement in automatic text classification for authorship attribution *Ingenio Magno, 6* **(2), 62-74.**

Antonio Rico-Sulayes
Grupo de Investigación en Lingüística Aplicada,
Universidad de las Américas Puebla,
antonio.rico@udlap.mx

Fechas de recepción: 14 de Junio de 2015
Fecha de aprobación: 18 de Diciembre de 2015

Abstract

In authorship attribution, the task of correctly assigning an anonymized document to an author within a predefined set of subjects, various measurements to evaluate classification systems have been used in the research literature. As will be discussed in this article, some of these measurements may differ diametrically. For research purposes, the evaluation of an automatic text classification system, such as the one that may be used for authorship attribution, may report a number of different performance measurements. However, some of the previously used figures are either too optimistic or lack generalizability. In addition to this issues, law-oriented research has pointed out the importance of having an error rate for the legal admissibility not only of this type of text classification task but of any piece of potential evidence in general. Considering the circumstances, the use of a single measurement in authorship attribution is proposed in this paper. Also, the implications of using this figure instead of others presented by researchers are discussed. At the same time, the importance of presenting this measurement along other relevant experimental settings, such as the number of categories (or authors in this context), is explained. The discussion is supported with the presentation of a set of authorship attribution experiments that utilize data from users of crime-related social media.

Keywords: classification systems, evaluation measurements, authorship attribution

Resumen

En la atribución de autoría, tarea que consiste en la asignación correcta de un documento anonimizado a un autor que es parte de un conjunto de sujetos, diversas medidas para la evaluación de sistemas clasificatorios han sido utilizadas por los investigadores del área. Como se argumenta en este artículo, algunas de estas medidas difieren diametralmente. Con fines investigativos, la evaluación de un sistema de clasificación automática de textos, como el que se puede llegar a utilizar en la atribución de autoría, puede reportar varias medidas diferentes sobre el desempeño del sistema; sin embargo, algunas de las figuras utilizadas previamente son, o bien demasiado optimistas, o bien poco generalizables. Además de estos problemas, la investigación en el ámbito legal ha enfatizado la importancia de contar con un índice de error para la aceptabilidad judicial no solo de este tipo de tarea de clasificación de texto, sino de cualquier prueba potencial en general. Por todo lo anterior, en este artículo se propone el uso de una medida única en la atribución de autoría. También se discuten las implicaciones de utilizar esta medida por encima de otras presentadas por algunos investigadores. Además, se expone la importancia de presentar esta medida en combinación con otras condiciones experimentales relevantes, como el número de categorías (o de autores en este contexto). La discusión se apoya de la presentación de una serie de experimentos de atribución de autoría que usan textos de usuarios de redes sociales relacionadas con el crimen.

Palabras clave: sistemas de clasificación, medidas de evaluación, atribución de autoría

Resumo

Na atribuição de autoria, uma tarefa que consiste na atribuição correta de um documento anônimo a um autor que faz parte de um conjunto de indivíduos, diversas medidas para a avaliação de sistemas de classificação tem sido usadas pelos pesquisadores da área. Conforme argumentado neste artigo, algumas destas medidas são diametralmente opostas. Para fins de investigação, a avaliação de um sistema de classificação automática de

textos, como o utilizado na atribuição de autoria, pode reportar várias medidas diferentes sobre o desempenho do sistema, porém, algumas das figuras utilizadas anteriormente são muito otimistas ou pouco generalizáveis. Além destes problemas, a pesquisa no âmbito legal tem enfatizado a importância de se ter uma taxa de erro para a aceitabilidade judicial não só deste tipo de tarefa de classificação de texto, mas qualquer evidência em geral. Por tudo o que foi citado anteriormente, este artigo propõe o uso de uma medida única na atribuição de autoria. Também são debatidas as implicações associadas à utilização desta medida acima das demais apresentadas por alguns pesquisadores. Além disso, se expõe a importância de apresentar esta medida em combinação com outras condições experimentais relevantes, tais como o número de categorias (ou autores neste contexto).
A discussão baseia-se na apresentação de uma série de experimentos de atribuição de autoria que utilizam os textos dos usuários de redes sociais relacionadas com o crime.

Palavras chave: sistemas de classificação, medidas de avaliação , atribuição de autoria.

1. Introduction

The need to identify the author of some given text is commonly present in legal contexts. McMenamin (2002), for example, gathers a comprehensive list of cases in the United States which required some sort of author identification for a written document. The conditions to perform this identification may vary significantly, giving origin to a wide range of related tasks (Rico-Sulayes, 2012). Among these tasks, authorship attribution represents an experimental design widely used for research purposes. This experimental design allows researchers to explore the possibility to respond to the legal need just mentioned. In authorship attribution, a system attempts to correctly assign an anonymized document to an author within a predefined set of subjects. Authorship attribution requires, then, a classification system. Given some document *x*, the system has to match this document to some individual in a set of potential authors {author *a*, author *b*, … author *n*}. This kind of exercise represents an idealized experimental setting, which allows researchers to obtain an error rate for their classification systems. Having an error rate is one of the desirable characteristics for any expert testimony to be admitted in court in the United States (Howald, 2008; Solan & Tiersma, 2004, 2005).

If reporting an error rate has this level of importance for the practical application of authorship attribution, one would expect its definition to be well established in the literature. However, as this task has benefited from the widespread availability of computer implemented classification algorithms, a number of different measurements have been used to report the success rate of authorship attribution experiments. Given this variation in the calculation of a success rate, there has also been a disparity in the calculation and presentation of the complementary figure for an error rate.

In this article, an analysis of a number of differing success rate figures in authorship attribution will be followed by a proposal to use a unique performance measurement. In order to argue for this proposal, experimental research results will be used. These results have been obtained conducting multiple authorship attribution tasks on users of crime-related social media.

2. Disparity in Authorship attribution performance measurements

In a classification task in general, a true positive (TP) represents a tested event correctly classified as belonging to its true class. Thus, in authorship attribution,

where the ultimate goal is to assign an anonymized document or text excerpt to its actual author among some given set of subjects, this correct assignment represents a TP. The proportion of TP obtained in a classification experiment is equivalent to the concept of *accuracy*.

A figure complementary to accuracy is the false positive proportion. In any classification task, a false positive (FP) represents an event incorrectly classified as belonging to a class other than its true class. For authorship attribution, then, an FP is given by a document or text excerpt incorrectly classified as belonging to an author other than its true author. The proportion of FP is equal to the *error rate*.

Table 1 below exemplifies a confusion matrix which is part of the results output commonly produced by classification software packages, such as *Weka* or *SPSS*. This table shows the accuracy or proportion of TP obtained for each of 10 categories in a classification task. This individual TP proportions are represented in the diagonal line formed by the cells with the same class in the x and y axes. Since in this example all classes have the same number of events or observations (ten), an average of these proportions is equal to the accuracy for the whole experiment.

Table 1. TP and FP proportion for classification with ten categories

Actual	Predicted class										
class	a	b	c	d	e	f	g	h	i	j	FP
a	0.9				0.1						0.1
b		0.5							0.5		0.5
c			1								0
d				1							0
e		0.4			0.1		0.5				0.9
f						1					0
g							1				0
h								0			0
i									1		0
j										1	0
										TP	ER
										0.85	0.15

In the context of authorship attribution, documents or text excerpts incorrectly classified as belonging to an author other than its true author are FP, the average proportion of FP per class is summarized in the right-most column of Table 1. Again, with a balanced number of events for all classes, an average of these proportions is equal to the error rate (ER) for an experiment as a whole. This figure is shown in the right bottom of the table. Since one event was incorrectly classified for class *a*, five for class *b*, and nine for class *e*, the total number of events incorrectly classified, fifteen, is divided by the total of events, 100. This renders an FP proportion or error rate of 0.15.

Understood as the proportion of TP, accuracy is the most commonly used performance measurement in information retrieval tasks (Manning, Raghavan & Schütze, 2008). When this figure is reported, the error rate is equivalent to the proportion of FP. However, in a comprehensive review of authorship attribution

studies, it is possible to find out that not all of them use the proportions of TP or FP as the central figures to evaluate their experiments. Surveying 33 authorship attribution studies, Rico-Sulayes (2012) identifies six studies that use either modifications of this measurement or completely separate, alternative figures. The rest of this part 2 will present and discuss these alternative performance measurements.

A. An optimistic true negative proportion

In authorship attribution research, a performance measurement that has been used to report experiment results is the proportion of true negatives (TN). A total of two, out of the six articles mentioned above, use this alternative success rate figure (Chaski, 2005, 2007). In the general context of classification, a TN represents a tested event correctly classified as not belonging to a class other than its true class. In the specific context of authorship attribution a TN is equal to the correct classification of a testing document or text excerpt as not belonging to an author other than its true author.

The problem associated with reporting the proportion of TN is that the success rate in this case tends to be too optimistic (Manning et al., 2008). For any problem that has more than two classes (or authors), the number of TN is much larger, compared to the number of TP. For example, in an authorship attribution experiment with documents by three authors, whenever the classification of a text renders one TP, it simultaneously gives two TN. This means that if a text is correctly assigned to its true author, it is tacitly not assigned to all the other subjects in the set of potential authors. Even if the assignment is wrong, the text is correctly not assigned to all other authors minus one.

The problem with this imbalance between TP and TN is that the number of TN grows at a very fast rate as the number of categories in the classification increases. Going back to the example in Table 1, with ten categories and ten events per category, a perfect classification exercise can produce a maximum number of 900 TN, compared to only 100 TP. This affects the proportion of both TP and TN because any number of negative assignments will have to be divided by this rather large number. As a result, the proportion of TN becomes very large and the proportion of false negatives (FN), the events incorrectly classified as not belonging to a class other than their true class, very small.

Table 2 shows the same results for the experiment presented in Table 1, but they are expressed now in terms of TN and FN. In the context of authorship attribution, FN represent documents or text excerpts incorrectly classified as not belonging to an author other than its true author. In Table 2, the diagonal line formed by the cells with the same class in the x and y axes shows the proportions for FN per class. The average of these proportions is equal to the proportion of FN for the whole experiment. The right-most column of Table 2 shows the proportion of TN for individual classes, as well as for the whole experiment at the right bottom.

Table 2. TN and FN proportion for classification with ten categories

Actual class	Predicted class										
	a	b	c	d	e	f	g	h	i	j	TN
a	0.01	1	1	1	0.9	1	1	1	1	1	0.99
b	1	0.06	1	1	1	1	1	1	0.5	1	0.94
c	1	1		1	1	1	1	1	1	1	0.9
d	1	1	1		1	1	1	1	1	1	1
e	1	0.6	1	1	0.1	1	0.5	1	1	1	1
f	1	1	1	1	1		1	1	1	1	1
g	1	1	1	1	1	1		1	1	1	1
h	1	1	1	1	1	1	1		1	1	1
i	1	1	1	1	1	1	1	1		1	1
j	1	1	1	1	1	1	1	1	1		1
										FN 0.017	0.983

As can be seen comparing Table 1 with Table 2, given the experimental conditions for the classification reported in these two tables, a TP proportion of 0.85 renders a TN proportion of 0.983, and a FP proportion of 0.15 gives a 0.017 FN proportion. Namely, if the report of results for some experiment is done in terms of TN or FN, rather than TP or FP, the classification system seems much more effective, even if the experiment and its results are exactly the same.

The extreme scenario for this bias is present when there are many categories or events, and a low accuracy. Large TN numbers can render near perfect proportion figures, even when there are no correct assignments of events to their classes. For example, in an authorship attribution experiment with fifty authors and ten documents by each author, a 0 accuracy or TP proportion will still render a TN proportion of 0.9796 (which results from dividing 24,000 correct negative rejections by 24,500 possible ones). This is at least the situation for classifiers that do not over generate TP. An example of this kind of classifier is the discriminant analysis, which has been extensively used in authorship attribution (Baayen et al., 2002; Chaski, 2005, 2007; Grant, 2007; Mikros & Argiri, 2007; Rico-Sulayes, 2011; Spassova, 2008, 2009; Spassova & Turell, 2007; Stamatatos, Fakotakis & Kokkinakis, 2001; Tambouratzis & Vassiliou, 2007).

B. Non-generalizable, non-cross validated results

Another issue in the evaluation of authorship attribution experiments is the use of non-cross validated results. Two authorship attribution studies out of the six mentioned at the beginning of part 2 offer results that were not cross validated (Grant, 2007; Spassova, 2009). Cross validation is a standard evaluation procedure in classification in general. It has also been extensively used in many tasks of computational linguistics (Jurafsky & Martin, 2008), the area to which automatic text classification belongs. Although there are various forms of cross validation, they all require dividing observation data into training and testing events.

For a classification experiment, the training events are used to create a statistical model of the different categories. Then, the testing events are classified through a comparison with the statistical models created from the training data. Whenever a testing event is classified, it is not part of the training data used to build the category models. The process of setting apart some

section of the observation data to classify it is repeated recursively until all the events are classified without ever using them to build the category models in the classification system.

As noted above, there are different forms of cross-validation. Two of the most common in classification tasks are leave-one-out cross validation and *n*-fold cross validation (Witten, Frank & Hall, 2011). In a leave-one-out cross validation, all events but one in a category are used as training data. The held out event is then classified comparing it to the rest of the training data. The process is repeated until all events are classified (Burns & Burns, 2008).

Another commonly used form of cross validation is *n*-fold cross validation. In an *n*-fold cross validation, all observation data is divided into *n* number of parts. The events in *n*-1 parts are used as the training data, and the events in the remaining *n*th part are classified based on the models created with the former data set. The process is repeated *n* times until all events are classified without having used any testing data during training (Jurafsky & Martin, 2008).

Cross-validation is used to mitigate one common problem in the evaluation of predictive models, the limited availability of testing data (Witten et al., 2011). It is important to use as much data as possible to make the predictive models efficient, but including the testing data in their construction exposes the classification system to the answers it is supposed to find out on its own, making the system biased.

Beyond the bias of knowing the results in advance, a different kind of problem, also related to the lack of a cross validated design, is the use of either a too small or a too large set of actual testing data. If one sets apart a small portion of the observation data, and uses it as the only test, there is the risk that this testing data is not representative. Presenting results from experiments with a small, non-representative testing data set may lead to wrong expectations about the performance of a given system once it is applied to new data. On the contrary, if we use a significant portion of the observation data, we may not have enough information in the training set to build an efficient model for all categories. Using cross validation, all available data is tested at some point, and a large portion of the available data is always utilized to create the statistical model for each category.

The specific issue with the lack of cross validated experiments in Grant (2007) and Spassova (2009) is the use of very small testing sets, from which is difficult to generalize any trend. Grant (2007) uses a very small testing set with three events only. He calls these events "query" cases. Spassova (2009) also uses non-cross validates results for twelve events, which she calls pseudo-anonymous cases. Some statistical programs allow the researcher to classify testing sets one by one, so the researcher can design manually what information gets in each *n*th part of a cross validation. However, in order to render generalizable results, researchers are expected to either use larger, representative testing data sets or continue with the classification of all *n*th parts. Failing to comply with any of these two requirements, the classification results lose their statistical generalizability.

C. A reduced set of authors

One last issue in the report of results for authorship attribution has to do with the manipulation and further selection of categories. In the context of authorship attribution, this implies a reduction of the set of authors in the experiment. With experiments in this task usually going from 2 to 100 authors (Rico-Sulayes, 2012), Koppel, Schler & Messeri (2008) and Koppel, Schler & Argamon (2009) attempt to perform a classification with a very large set of 10,000 authors. To tackle this task, these researchers devise a statistical technique to identify documents whose authors are especially

difficult to distinguish from others because their use of classificatory features is very similar. After identifying and eliminating a very large portion of the categories most difficult to classify (namely, the authors whose writing is especially hard to discriminate), the researchers' accuracy or TP proportion increases significantly.

In order to interpret the results obtained through this method of elimination of difficult categories, it is important to contextualize authorship attribution as a biometric test. Although this task is not traditionally included in biometrics manuals (Petrovska-Delacretaz, Chollet & Dorizzi, 2009), its goal, identifying the author of an anonymous text by comparing the writing samples of a set of individuals, matches the definition of a biometric process (Zvetco Biometrics, 2012). Regarding biometric processes, Bolle et al. (2004) note that the elimination of poor data, usually events or observations, may be an acceptable quality control practice in the construction of biometric databases, but it can make the evaluation of a classification system "look arbitrarily good" (p. 166). Besides this problem, an even bigger issue is raised when not only are biometric events eliminated by means of substituting them with better quality events (as when a fingerprint or picture is retaken), but there is actually an elimination of biometric subjects or categories from the observation data.

Applying progressively their technique to eliminate difficult categories, or authors hard to distinguish from others, Koppel et al. (2008) achieve a TP proportion of 0.9 for 30% of the subjects in their original database. A similar result is obtained in Koppel et al. (2009), where the scholars achieve a TP proportion of 0.882 for 31.3% of all the subjects in their full database. In the first study, the researchers argue that their technique is equivalent to an "I don't know" comment or expert opinion in a law enforcement scenario. However, they do not report any legal cases or consultation where their method has been used and accepted. Despite of their admissibility in the legal context, which is not argued for by the researchers, the results in these studies represent only a fraction of all authors in the whole database. Considering all the original data, the actual TP proportion for the whole data base of authors in Koppel et al. (2008) is 0.27 and 0.276 in Koppel et al. (2009). Therefore, without a reduced set of categories, the overall TP proportion in the classification of the entire data set looks much less effective.

3. Application of a single measurement under relevant experimental conditions

The lack of consistency in the presentation of results in authorship attribution, as has been exemplified in part 2, represents a major drawback for the practical application of this biometric task. As commented in part 1, presenting an error rate can be a decisive characteristic for the legal admissibility of expert testimony in countries like the United States (Howald, 2008; Solan & Tiersma, 2004, 2005). For this reason, this article proposes the use of a single, unified success rate figure, accuracy or the proportion of TPs, to report experiments results in this task. Its complementary figure, the proportion of FP, is proposed as a standardized error rate. As to the three alternatives for the evaluation of classifications systems discussed above, the proportion of TN (section 2-A), the use of non-cross validated results (section 2-B), and the optimization of the observation data through the elimination of categories (section 2-C), they all posit significant problems.

The most important weaknesses in these three alternative evaluation frameworks are, for the first and third alternatives, the excessive optimism of the reported figures, and for the second alternative, the lack of generalizability of the estimated performance for the classification system. In face of these problems, the TP proportion represents a more stringent, but realistic approach. In this sense, the author of this article believes that instead of devising new evaluation

systems, it is more useful to present the formerly proposed measurement (TP proportions) along other relevant experimental settings, such as the number of categories in the classification task.

Although the number of categories in authorship attribution is usually presented by researchers (all 33 studies surveyed in Rico-Sulayes, 2012, do include this piece of information), exploring the effect of the number of categories in the classification as it progressively increases or decreases, is not very common in the literature. Only two out of the 33 studies mentioned present the success rate of their experiments along four or more intervals in the number of categories (Grieve, 2007; Zheng et al., 2006).

This piece of information, a gradually changing number of categories, is important because this experimental condition has been shown to influence the success rate of the classification. In both, Grieve (2007) and Zheng et al. (2006), as well as in another study that uses less than four intervals for the number of categories (Argamon, Šarić& Stein, 2003), it has been shown that there is a decrease in the success rate of authorship attribution experiments as the number of subjects included in the classification task increases. In order to show the consistency and appropriateness of presenting the TP rate along several intervals for the number of categories in the task, this article presents the results in a series of very successful authorship attribution experiments with up to 40 categories.

A. Experimental data

The data for the following experiments has been harvested from what was one of the first online forums devoted to the topic of organized crime in Mexico (Foros Blog del Narco, 2010). With a complex history that has resulted in the murder of some of their users, this type of social media appeared in this country in 2010. The online forum retrieved here was created in April of that year. Six months after its creation, when it had become popular in the media (Rico-Sulayes, 2011), 41,751 users' contributions were spidered from the forum. Once this collected data was processed and cleaned, 37,571 contributions posted by 1,026 logged users with a total 2,128,049 word tokens were used to build 39 data sets. With an increasing number of categories, from two to forty authors, these data sets included contributions from prolific users. Users were considered prolific if they had posted a minimum of 40 contributions, with at least 2,000 words of original text.

B. Classification method

In order to perform an authorship attribution task in the 39 data sets described in section 2-A, the most commonly used classification algorithms in this task (the decision tree C4.5, discriminant analysis, multivariate naïve Bayes, the Bernoulli model of naïve Bayes, and support vector machines) were tested, along a number of classificatory feature sets (Rico-Sulayes, 2014). These various feature sets were built from an original list of features that included nineteen structural features (such as the use of font colors, emoticons, and hyperlinks), 132 syntactical features (sequences of function words, mostly prepositions and conjunctions), and up to 13,098 lexical features (including primarily word forms or types). The different data sets tested were produced using three stochastic techniques to select discriminatory features (the two most common techniques in the task, plain frequency and information gain, as well as correlation-based feature subset selection). With a total of 780 experiments, this article reports the results for the most successful combination of a classifier and a feature set over the 39 data sets.

C. Result analysis

The best averaged result over all data sets (0.947) was obtained by the multivariate naïve Bayes classifier in combination with a feature set stochastically reduced using an information gain score greater than zero.

Figure 1 below shows the TP proportion obtained by this classification method in 20 different intervals for the experiments with an ever number of categories, from two to 40 authors.

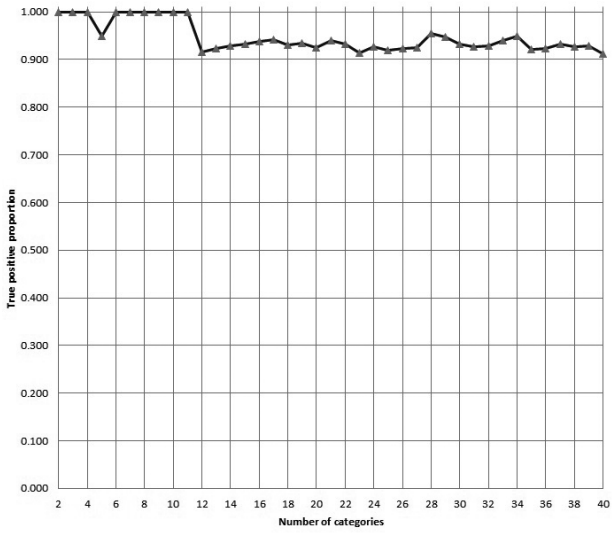

Figure 1. TP proportion at twenty category intervals

Although the most successful classification method remains fairly steady in its results with the different data sets, a trend in its accuracy to decrease as the number of categories increases can be appreciated in Fig. 1. This is consistent with the research in the area, mentioned at the beginning of this part 2 (Argamon et al., 2003; Grieve, 2007; Zheng et al., 2006). Besides, this tendency can be also observed in the rest of the experiments with the other classifiers and feature sets mentioned in section 3-B. This is an example of how using a unified performance measurement makes research efforts comparable and consistent. This comparison may be difficult and the consistency simply unattainable when alternative evaluation frameworks are used.

D. Discussion

Besides the just mentioned number of categories in the classification, there are other variables whose effect has been studied by researchers. An example of this is the effect of the size of the training data set in the classification performance (Burrows, 2002; Peng, Schuurmans, Keselj & Wang, 2003; Stamatatos et al., 2001; Zheng et al., 2006). Consistent in the four studies just cited, a trend has been found in the TP proportion to improve as the amount of training data increases, in terms of either words (Burrows, 2002) or documents (Peng et al., 2003; Stamatatos et al., 2001; Zheng et al., 2006).

What is interesting regarding this variable is that there is one more study that explores the effect of the amount of training data in the classification (Grant, 2007), but it finds the opposite trend, a decrease in accuracy as the amount of data in the training set increases. Although the author of this article does not compare his results to the four previous studies exploring this variable effect, it is worth noting that Grant's article is one of the six studies reviewed in section 2-B, and whose lack of generalizability due to its non-cross validated designed was pointed out at the beginning of this part 2.

4. Conclusions

This article has identified a lack of consistency in the report of the success rate and the complimentary error rate by authorship attribution researchers. The importance of these two measurements has been identified, as stated in the legal literature that discusses the practical implications of this text classification task. Given this importance, the disparity of performance measurements has been discussed and a single evaluation measurement has been proposed.

The consistency that can be achieved when using a unique measurement has been shown through the presentation of a wide range of experiments and the comparison of experiments with the best results to results in previous research. In this sense, it has also been shown how the studies that do not follow the standard evaluation here argued for may go astray in their findings and differ from what the rest of the related research shows.

Referencias

Argamon, S., Šari, M. & Stein, S. S. (2003). Style mining of electronic messages for multiple authorship discrimination: first results. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Baayen, H., van Halteren, H., Neijt, A. & Tweedie, F. (2002). An experiment in authorship attribution (pp. 29-37). *Proceedings of JADT 2002: Sixth International Conference on Textual Data Statistical Analysis*.

Bolle, R. M., Connell, J. H., Pankanti, S., Ratha, N. K. y Senior, A. W. (2004). *Guide to biometrics*. New York: Springer-Verlag.

Burns, R. B. & Burns, R. A. (2008). *Business research methods and statistics using SPSS*. UK: Sage.

Burrows, J. (2002). Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267-86.

Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence, 4*(1), 1-13.

Chaski, C. E. (2007). The keyboard dilemma and authorship identification. In P. Craiger & S. Shenoi (Eds.), *Advances in Digital Forensics III* (pp. 133-146). New York: Springer.

Foros Blog del Narco. (2010). Retrieved from http://www.foro.blogdelnarco.com/

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language and the Law, 14*(1), 1-25.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 425-442.

Howald, B. S. (2008). Authorship attribution under the rules of evidence: empirical approaches - a Layperson's Legal System. *International Journal of Speech, Language and the Law, 15*(2), 219-247.

Jurafsky, D. & Martin, J. H. (2008). *Speech and language processing: an introduction to language natural processing, computational linguistics, and speech recognition* (2nd ed.). Upper-Saddle River: Pearson-Prentice Hall.

Manning, C. D., Raghavan, P. y Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge.

McMenamin, G. R. (2002). *Forensic linguistics: advances in forensic stylistics*. Boca Raton: CRC.

Mikros, G. K. & Argiri, E. K. (2007). Investigating topic influence in authorship attribution. In B. Stein, M. Koppel & E. Stamatatos (Eds.), *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007*.

Koppel , M., Schler, J., & Argamon, S. (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology, 60*(1), 9-26.

Koppel, M., Schler, J., & Messeri, E. (2008). Authorship Attribution in Law Enforcement Scenarios. In C.S. Gal, P. Kantor, & B. Saphira (Eds.), *Security Informatics and Terrorism: Patrolling the Web* (pp.111-119). Amsterdam: IOS.

Peng, F., Schuurmans, D., Keselj, V. & Wang, S. (2003). Language independent authorship attribution using character level language models. *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics: Vol. 1* (pp. 267-274). Stroudsburg: Association for Computational Linguistics.

Petrovska-Delacretaz, D., Chollet, G. y Dorizzi, B. (2009). *Guide to biometric reference systems and performance evaluation*. London: Springer-Verlag.

Rico-Sulayes, A. (2011). Statistical authorship attribution of Mexican drug trafficking online forum posts. *International Journal of Speech, Language and the Law, 18*(1), 53-74.

Rico-Sulayes, A. (2012). Quantitative authorship attribution of users of Mexican drug dealing related online forums (PhD dissertation, Georgetown University). Retrieved from https://repository.library.georgetown.edu/handle/10822/557726

Rico-Sulayes, A. (2014). Técnicas de reducción, algoritmos resistentes al ruido o ambos. Opciones para el manejo de rasgos clasificatorios en la atribución de autoría. *Research in Computing Science, 80*.

Solan, L. M. & Tiersma, P. M. (2004). Author Identification in American Courts. *Applied Linguistics, 25*(4), 448-465.

Solan, L. M. & Tiersma, P. M. (2005). *Speaking of Crime: The Language of Criminal Justice*. Chicago: University of Chicago.

Spassova, M. S. (2008). Las perífrasis verbales del español en la atribución forense de autoría. In R. Monroy & A. Sánchez (Eds.), *25 años de lingüística en España: hitos y retos. Actas del XXVI Congreso de AESLA* (pp. 605-614). Murcia: Universidad de Murcia.

Spassova, M. S. (2009). *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español* (PhD dissertation, Universitat Pompeu Fabra, Barcelona). Retrieved from http://repositori.upf.edu/handle/10230/12285

Spassova, M. S. & Turell, M. T. (2007). The use of morphosyntactically annotated tag sequences as markers of authorship. In M. T. Turell, J. Cicres, and M. S. Spassova (Eds.), *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006* (pp. 229-237). Barcelona: Documenta Universitaria.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities, 35*, 193-214.

Tambouratzis, G. & Vassiliou, M. (2007). Employing thematic variables for enhancing classification accuracy within author discrimination experiments. *Literary and Linguistic Computing, 22*(2), 207-224.

Witten, I. H., Frank, E. & Hall, M. A. (2011). *Data mining: practical machine learning tools and techiniques* (3rd ed.). Burlington: Morgan Kaufmann.

Zheng, R., Li, J., Chen, H. & 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.

Zvetco Biometrics. (2012). *Biometric Knowledge Center*. Retrieved from http://www.zvetcobiometrics.com/Support/definitions.jsp