Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 175 (2020) 671–676

Procedia Computer Science

www.elsevier.com/locate/procedia

The 7th International Symposium on Emerging Inter-networks, Communication and Mobility (EICM)
August 9-12, 2020, Leuven, Belgium

Classification of authors for an automatic recommendation process for criminal responsibility

Amelec Viloria [a,*], Omar Bonerge Pineda Lezama[b], Eduardo Chang[c]

[a,b] *Universidad de la Costa, Barranquilla, Colombia.*
[c] *Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras*

Abstract

One problem in classifying tasks is the handling of features that characterize classes. When the list of features is long, a noise resistant algorithm of irrelevant features can be used, or these features can be reduced. Authorship attribution is a task that assigns an anonymous text to a subject on a list of possible authors, has been widely addressed as an automatic text classification task. In it, n-grams can produce long lists of features even in small corpora. Despite this, there is a lack of research exposing the effects of using noise-resistant algorithms, reducing traits, or combining both options. This paper responds to this lack by using contributions to discussion forums related to organized crime. The results show that the classifiers evaluated, in general, benefit from feature reduction, and that, thanks to such reduction, even classical algorithms outperform state-of-the-art classifiers considered highly noise resistant.

Keywords: Authorship attribution; Classification features; Noise resistant algorithms; Feature reduction..

* Corresponding author. Tel.: +57-3046238313.
E-mail address: aviloria7@cuc.edu.co

10.1016/j.procs.2020.07.098

1. Introduction

If authorship attribution is defined as the assignment of an anonymous text to a subject within a list of possible authors, this task constitutes a text classification problem. However, if this problem is tackled using automated methods, the problem lies with automatic text classification, an area served by information retrieval. As a problem of automatic text classification, authorship attribution uses two primary elements. On the one hand, it requires a selection of classifying features, which discriminate elements from different classes. On the other hand, attribution of authorship uses a classification method that processes the traits. In this context, the classification method is used to attribute a certain text to a specific subject [1].

With respect to the first essential element for attribution of authorship, the selection of classifying features, researchers dedicated to this task have been proposing new features consistently for several decades. At the end of the last century, more than 1,000 different features were identified in more than 300 papers devoted to this classificatory task [2]. This number of features has increased dramatically in recent years due to the introduction of textual features that are automatically labeled, such as n-grams. N-grams easily produce lists of several thousand elements even in relatively small corpora (collections of natural language texts), such as those typically used in authorship attribution [3].

In contrast to the proliferation of authorship features, recent research findings suggest that the selection of authorship traits is the primary element in improving outcomes for this task [4]. According to the above-mentioned research, this selection is even more important than the development of classification algorithms. In text classification in general, the motivation for using reduced lists of traits is that highly discriminatory traits are more efficient and obtain a higher precision in the results [5]. These discriminatory features avoid the noise of long lists, which include redundant or non-discriminatory features. This type of feature is particularly inefficient when applied to new data sets. One response to this problem in authoring attribution is advanced text classification algorithms, such as support vector machines (SVM), which can compensate for the noise of long lists of features [6], [7]. The second answer is the use of techniques for reducing such lists, such as selecting features with higher frequency or high rates of mutual information. This solution has also been widely used in the attribution of authorship [8]. Against these two possibilities in the management of classifying features for attribution of authorship (the use of algorithms that compensate for the noise of non-discriminatory or redundant features and the alternative reduction of feature lists with resources external to the classifier), the literature has not compared the results of the two options.

This article responds to this shortcoming by comparing the most common methods of classification and reduction of features in the attribution of authorship. In addition, this paper introduces a feature reduction method never before used in this task. In evaluating the different combinations of classifiers and features reduction techniques, this paper uses data from social media related to organized crime in Colombia. The paper concludes by showing that classification methods with a long tradition of attribution of authorship [1] can be combined with feature reduction techniques (both known and new techniques in this context) and that these combinations equal and exceed the results obtained by state-of-the-art classifiers.

2. Social media and organized crime

This paper includes user contributions published on one of the first sites related to organized crime in Colombia. This site, created in April 2002, originally hosted a discussion forum dedicated to this topic [9]. The contributions to this forum were recovered by copying all the messages published during the first half year of the forum's life. This allowed for the recovery of 58,254 messages published in 6,487 conversations. After debugging all the recovered messages (deleting copies and messages from anonymous users) 48,784 messages created by registered users of the forum were identified. These messages belong to 2,562 different users and contain a total of 3,478,254 instances of words or tokens.

2.1. Experimental data

With the data retrieved from the discussion forum mentioned above, several corpora were created to explore the effects of feature reduction in combination with various classifiers, common in the attribution of authorship. Among the 2,562 users who produced messages using a user account, those who had a minimum of 40 individual messages with a minimum of 2,000 words of original text in the sum of all their messages were selected. Using these two selection criteria, 254 forum users were identified as meeting these criteria.

The minimum number of words of original text required to select forum users (2,000 words) is at the lower end of what previous studies have used in the attribution of authorship. For example, among the researchers reporting this experimental data some have used 2,000, 8,000, 15,000, 33,000, 40,000, and 55,000 words, [10], [11], [12], respectively. This amount of text is used as training data to represent each subject in the pool of potential authors during classification. The other criterion of selection of subjects as potential authors (a minimum of 40 messages), was used to rule out sporadic users with few messages of a certain length. By randomly sampling the 106 selected, 40 users were identified with which 39 corpora were constructed. In these corpora, the number of subjects ranges from a minimum of 2 to a maximum of 40.

Dividing the 2,000 words of original text from each subject, 4 sub-samples of approximately 500 words each were constructed. These sub-samples were constructed by randomly adding messages, from among all the contributions of each of the forum users, preserving the integrity of the individual messages. Each of these sub-samples was used as a unit in the test data used for the ranking. The range in size of these test sub-samples (478-541 words) is also at the lower end of what previous studies using integrity messages have employed.

3. Attribution of authorship as automatic text classification

The attribution of authorship has been widely addressed, both by information retrieval researchers and by forensic linguistics and stylistics [13].

3.1. Authorship Qualifying Features

As for the features used to carry out the classification, a previous selection of lexical, syntactic and structural features was taken as a starting point. The lexical features included a list of all word unigrams (equivalent to all types or differentiated lexical forms). As this list is dependent on the corpus from which it is extracted, the size of the list ranged from 1,402 types for the smallest corpus, to 13,089 for the largest corpus. It should be noted that punctuation was removed from the lexical units to which it was attached and the separate signs were used as independent lexical unigrams, a common procedure in the attribution of authorship [14] and [15]. As for syntactic features, a previously collected list (for another classification task) of functional lexical elements with pluriverbals, i.e. more than one word, was used. These elements are mainly composed of a preposition plus other lexical elements, such as 'after(l)' or 'away from(l)', or of a conjunction combined with other words, such as 'after' or 'while'. The default list of pluriverbal functional lexical elements, whose instances were labeled in the corpora, had 132 elements in total, with 68 bigrams, 56 trigrams, and 7 tetragrams. Finally, the structural features were given by a pre-selected list of 19 elements, several of them previously used by the author of this paper [6]. These features include various text formatting features such as the use of underlines, bold, images, colors, and special font sizes. These features also include elements of electronic communications, such as the use of active and inactive hyperlinks, emoticons, in images and represented by keyboard characters, and the excessive reduplication of punctuation marks, as is often the case with exclamation marks.

3.2. Classification algorithms in the attribution of authorship

Research on attribution of authorship has employed a variety of classification algorithms. In a comprehensive review of 32 dedicated papers published in the last decade, 23 different classification algorithms were identified [1]. Although many of the identified classifiers appear only once in the literature, some algorithms have been used in several investigations. These algorithms include different implementations of the decision tree, C4.5, various forms of Bayesian analysis (such as multivariate and Bernoulli's model), different types of neural networks (such as artificial and so-called backpropagation neural networks [9]) and SVM. Also common in the attribution of authorship are some statistical classifiers, such as discriminant analysis (DA) and classifiers based on the Chi-square test. In addition, it should be mentioned that 10 of the 32 papers reviewed in the above-mentioned study use more than one classifier algorithm and compare the results obtained by the different selected algorithms.

As for the classifiers tested in this study, the 4 algorithms that have given the best results in the attribution of authorship were chosen, according to the exhaustive bibliographic review mentioned above [10]. The 4 classifiers are DA, multivariate Bayesian analysis (MBA), Bernoulli's Bayesian analysis (BBA) and SVM. In addition to these classifiers, the most commonly used baseline algorithm in this task, the C4.5 decision tree, was added in its implementation for Weka, J4.8.

3.3. Trait reduction techniques in the attribution of authorship

The study mentioned in the previous section [10] also reports an abundant use of trait reduction techniques in authorship research. Of the 32 studies reviewed, 17 were identified by their use of some feature reduction technique, or some method of feature assessment that allows for overall reduction. With less variation than the use of classifiers, the feature reduction techniques identified include information gain (IG), frequency (relative, absolute, or standardized), principal component analysis, some general statistical feature assessment methods (analysis of variance, ANOVA, analysis of covariance, ANCOVA, and two-way ANOVA), and two stepwise methods, Mahalanobis distance and Wilks' Lambda. One study also uses a list of empty words to remove them from its full list of features. Although only two studies compare more than one reduction technique, the use of two of them is noticeably more common than the others: frequency, used in six studies, and IG, in three.

In this research, the two most common reduction techniques, frequency and IG, were chosen, which are also those with which the best results have been reported in comparative studies. The frequency here was expressed as an absolute frequency with a minimum of 4 instances, a number equal to the number of sub-samples per author. In addition, it was decided to include a new reduction technique in the attribution of authorship, the so-called "correlation-based feature subset selection" (CFS). This reduction technique, first described in [7], was included because it was designed with the explicit intention of improving the performance of algorithms based on Bayesian analysis. As mentioned above, two versions of these algorithms, MBA and BBA, were used here.

4. Results

For the accuracy reported below, it represents the proportion of true positives or correct assignments of test sub-samples to their true authors. To obtain this precision in individual experiments, a classifier and a list of features (reduced or without reduction) were applied to the assignment of all test sub-samples to their respective authors. In addition, the calculation of the precision in the individual experiments was carried out by means of a cross validation design. The figures presented below in Table 1 represent the average precision obtained by the combination of a classifier and a list of traits when applied gradually to the 39 corpora.

Table 1. Average results of the classification in the 39 corpora

Reduction Technique	**Classifier**				
	C4.5	AD	MBA	BBA	**SVM**
None	0.465	0.324	0.755	0.847	**0.501**
Frequency	0.494	0.477	0.958	0.802	**0.835**
IG	0.688	0.711	0.936	0.854	**0.748**
CFS	**0.674**	**0.798**	**0.957**	**0.814**	**0.744**

5. Conclusions

With the global perspective provided by Table 1, it is possible to return to the list of options for the management of qualifying features mentioned in the title of this paper. From that list, the question arises: What is the best option to handle long lists of features in the attribution of authorship: to apply feature reduction techniques, to use noise resistant classification algorithms or to combine these two options? Regarding the first answer to the question, the application of feature reduction techniques, Table 1 shows that all classifiers clearly benefit from such application. At the same time, there are three trends in the benefit obtained through the various reduction techniques. On the one hand, two classifiers, C4.5 and DA, show a moderate improvement with the list reduced by frequency. The improvement ranges from 0.465 to 0.494 for C4.5 and from 0.324 to 0.477 for AD. However, the improvement in accuracy of these two classifiers is greater with techniques that require more elaborate calculation and probably more importantly produce much shorter lists of features. In this regard, it should be mentioned that the number of traits in the reduced list with the frequency criterion ranges from 229 in the corpus with two authors to 2,168 in the corpus with 40 authors, while IG and CFS generate considerably smaller lists, with ranges of 23-75 features and 7-27 features, respectively. The second trend observed in the application of reduction techniques is given by two other classifiers, MBA and SVM. These algorithms benefit significantly from all feature reduction techniques.

Figure 1 shows the accuracy obtained by combining the list of features selected by IG and MBA. This combination obtained the highest average accuracy over all the corpus. In the same figure, the accuracy obtained by MBA without any reduction of feature was included.

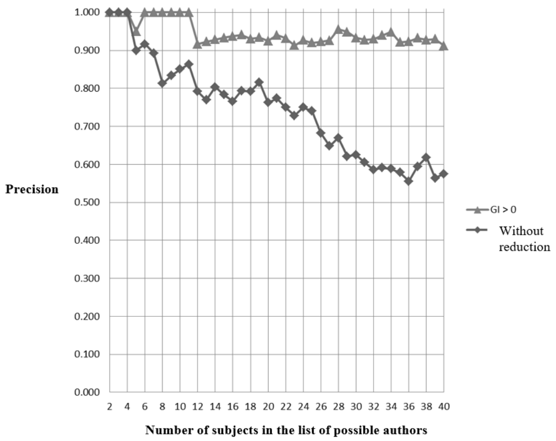

Fig. 1. Effect of feature reduction by IG with MBA

The top line in Figure 1, with vertices highlighted with triangles, corresponds to the accuracy of the MBA with the reduced list with IG, while the bottom line, with vertices highlighted with diamonds, describes the accuracy of the MBA when using the unreduced list. These two lines show the clear positive effect of feature reduction on the most successful combination of a reduction technique and a classifier in this study. As seen in the top line, the application of the feature reduction technique by IG allows the MBA classifier to maintain a constant accuracy above 0.900 throughout all the experiments with the 39 corpora, which include, as shown in the figure, from two to 40 authors.

Regarding the second option to handle classifying features, the use of noise-resistant classifiers, this work evaluated a state-of-the-art algorithm in machine learning, the SVM. This algorithm is considered to be particularly noise-resistant in the context of authorship attribution [3], [9]. In this study, SVM showed, on the one hand, that they are indeed highly resistant to noise, since they obtain their best average result with the reduced lists with the frequency criterion, which are comparatively long. However, the combination of a traditional classifier model, the MBA, in combination with any of the three feature reduction methods (frequency, IG and CFS), has the ability to overcome the best results obtained by SVM. The superiority of MBA, in its best performance achieved with IG, over the best performance of SVM, when combined with the frequency-reduced list, can be seen in Figure 2.

The third option for handling long lists of features is the simultaneous use of highly noise-resistant algorithms and reduction techniques. Whether or not this option should be used is clear from the last two points discussed. On the one hand, all classifiers benefited from feature reduction techniques. On the other hand, the highly noise-resistant algorithm showed the greatest benefit when using comparatively long reduced lists. However, it was the combination of a classifier that is not considered particularly noise-resistant and a noticeably reduced list (and debugging of those features that insert noise) that has obtained the best results in the evaluation of the multiple combinations of classifiers and reduction techniques. This third option has therefore been overcome, like the second option discussed above.

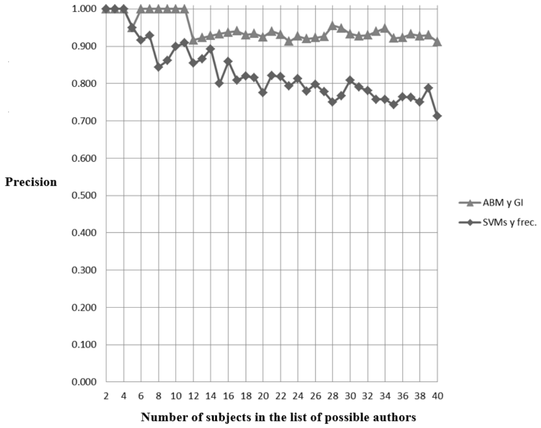

Fig. 2. Highest average accuracy obtained by MBA and SVM

References

[1] Vorobeva, A. A. (2016, April). Examining the performance of classification algorithms for imbalanced data sets in web author identification. In 2016 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT) (pp. 385-390). IEEE.

[2] Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., ... & Stamatatos, E. (2016). Authorship attribution for social media forensics. IEEE Transactions on Information Forensics and Security, 12(1), 5-33.

[3] Rico-Sulayes, A. (2017). Reducing Vector Space Dimensionality in Automatic Classification for Authorship Attribution. Revista Científica de Ingeniería Electrónica, Automática y Comunicaciones, 38(3), 26-35.

[4] Win, K. N., Li, K., Chen, J., Viger, P. F., & Li, K. (2019). Fingerprint classification and identification algorithms for criminal investigation: A survey. Future Generation Computer Systems.

[5] Tarmizi, N., Saee, S., & Ibrahim, D. H. A. (2020). Author identification for under-resourced language Kadazandusun. Indonesian Journal of Electrical Engineering and Computer Science, 17(1), 248-255.

[6] Sun, S. (2019). Application of Fuzzy Image Restoration in Criminal Investigation. Journal of Visual Communication and Image Representation, 102704.

[7] Boenninghoff, B., Nickel, R. M., Zeiler, S., & Kolossa, D. (2019, May). Similarity learning for authorship verification in social media. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2457-2461). IEEE.

[8] Watson, D. (2019). Source Code Stylometry and Authorship Attribution for Open Source (Master's thesis, University of Waterloo).

[9] Juola, P., Milička, J., & Zemánek, P. (2018). Authorship and time attribution of Arabic texts using JGAAP. In Intelligent Natural Language Processing: Trends and Applications (pp. 325-349). Springer, Cham.

[10] Hannah-Moffat, K. (2019). Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates. Theoretical Criminology, 23(4), 453-470.

[11] Mutanen, T. P., Metsomaa, J., Liljander, S., & Ilmoniemi, R. J. (2018). Automatic and robust noise suppression in EEG and MEG: The SOUND algorithm. Neuroimage, 166, 135-151.

[12] Usha, A., & Thampi, S. M. (2017, December). Authorship Analysis of Social Media Contents Using Tone and Personality Features. In International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage (pp. 212-228). Springer, Cham.

[13] Hasanov, A., & Mukanova, B. (2017). Fourier Collocation Algorithm for identification of a spacewise dependent source in wave equation from Neumann-type measured data. Applied Numerical Mathematics, 111, 49-63.

[14] Reddy, T. R., Vardhan, B. V., & Reddy, P. V. (2016). A survey on authorship profiling techniques. International Journal of Applied Engineering Research, 11(5), 3092-3102.

[15] Sun, F., Gu, Y., Cao, Y., Lu, Q., Bai, Y., Li, L., ... & Li, T. (2019). Novel flexible pressure sensor combining with dynamic-time-warping algorithm for handwriting identification. Sensors and Actuators A: Physical, 293, 70-76.