ISSN 1870-4069

Selection of Features for Attribution of Authorship Using a Genetic Algorithm and Support Vector Machine as a Function of Aptitude

Omar González Brito[1], José Luis Tapia Fabela[1], Silvia Salas Hernández[2]

[1] Universidad Autónoma del Estado de México, Unidad Académica Profesional Tianguistenco, Mexico

[2] Universidad Autónoma del Estado de México, Centro Universitario UAEM Atlacomulco, Mexico

{gonzalezbritoomar, joseluis.fabela, salashernandezsilvia}@gmail.com

Abstract. In the present investigation, a genetic algorithm was developed with a support vector machine as an aptitude function. This algorithm has the purpose of searching vector sub-spaces of features that improve the authorship attribution. This method reduces dimensionality, improves classification, and identifies which are the most significant features to differentiate between authors. To determine the author, it is necessary to analyze linguistic features or textual features that allow us to find the author's writing style. In the experiments carried out, it is observed that the results obtained are satisfactory, where they proposed six types of samples, of which three were balanced and three were imbalanced, where the samples of size five improved the accuracy by 14.0%, in the sample of size 10 the accuracy is improved by 5.3% and in the sample size 10:20 the accuracy is improved by 6.15%.

Keywords: Genetic algorithm, selection of features, support vector machine.

1 Introduction

At present, authorship analysis has become a major problem in many areas; among them, information retrieval, computational linguistics, and forensic linguistics. A great diversity of corpus has been created, covering different contexts such as; emails, journalistic notes, literary works, and social networks [1, 2, 3, 4, 5, 6]. The authorship attribution consists of identifying the author of a text within a set of candidates through their textual features that allow differentiating the writing style of an author [7, 8]. The computational problem of authorship attribution has been addressed mainly through the classification of texts [1, 8, 9, 10, 11].

In [12], it is shown that every classification task tends to have many features that can be relevant, irrelevant, and redundant. Irrelevant and redundant features degrade the performance of a classifier. This is sensitive to the selection of its features. When making the selection of features, the performance of the classifier is optimized [13], [14]. The lexical features mainly used in authorship attribution are the bag of words and n-grams models. Others found within the lexicons are the original words and the empty words, the count of the length of the sentences or the number of words in the text, etc. [15, 16, 17, 18].

The main advantage of these features is domain and context independence except in languages such as Chinese. These types of features do not require any processing to obtain them or the application of pre-processing for their analysis. The selection of features is intended to eliminate irrelevant, redundant features. For the selection of features, filter methods (statistical) and wrapper methods (learning) have been implemented. This makes it possible to eliminate unnecessary features and improve the classification [19]. The filter method selects the features utilizing a relevance criterion. This criterion can be the measure of distance or dependence.

The filter methods evaluate the features independently concerning the classes of the training set. The wrapper method selects subsets of features from a classifier. The criteria for selection are according to the learning algorithm, mainly search algorithms are implemented [19, 20, 21, 22]. When characterizing a text with the n-gram model, the dimensionality of the features is very high, within this set there are relevant, irrelevant, and redundant features [12].

To choose the relevant features of an author, a genetic algorithm was implemented as a method of selection of features, which has as its main objective the search for optimal vector subspaces, which allows improving the classification [23]. In the present investigation, they were obtained better results in three out of six samples and competitive results in the other samples compared with one of the works most referenced by the state of the art present in [11].

2 Related Works

Genetic algorithms are considered meta-heuristic methods, which mimic the process of natural selection proposed by Darwin. These adaptive methods arise with the need to find solutions to problems in a complex search space. The individuals with the best performance will have the possibility of being selected to pass to the next generation and reproduce these are considered possible solutions to a particular problem [24, 25, 26, 27].

In the literature review, different works have been found for the selection of features, one of them is presented in [23], where they implement a genetic algorithm and the k-closest neighbor's classifier; select the most significant words (features) to improve accuracy in the authorship of science fiction texts. Where the dimensionality to be treated is lower compared to the analysis of n-grams at the character level. The corpus used consisted of 503 science fiction text files in English with 17 authors, using 350 for the training phase and 153 for validation.

However, in the literature review for the text classification task, better results are observed by implementing a support vector machine. Another work is presented in [14],

where he implements a genetic algorithm with a support vector machine, used for the selection of the most significant features, applied to the detection of cancer with good results. The dimensionality of the features to be selected is lower and the classes are two (it has cancer and it does not have cancer), these considerably reduce the number of features obtaining good results with only nine features.

The corpus consisted of 200 images for training and 77 for validation, the results obtained exceed 70% accuracy. The parameters used for his genetic algorithm were roulette selection, two-point crossing, Gaussian mutation, with a population of 100 individuals, and 10 generations with a mutation probability of 0.01%.

On the other hand, in [28] they implement a genetic algorithm for the selection of features, it mentions that an important part within the classification of texts is the selection of features, this type of problem seeks to optimize the process of selection of features, due to this need to implement a genetic algorithm, this type of method allows finding the best solutions to particular problems.

In this work they implement a genetic algorithm that meets the following conditions; the set of features chosen by the algorithm had to represent the category of texts, the aptitude of the individuals was evaluated individually to ensure optimal similarity between individuals, the dimension of the vector of features had to be the smallest possible size, for its evaluation used cosine similarity metrics using an elitist selection operator, generating a random population of 100 individuals and 400 generations, with a mutation rate of 0.005%.

The Corpus consisted of 1200 documents from the SOGU laboratory, of which 800 were used for training and 400 for validation, taking into account 8 categories, the results obtained show that the genetic algorithm obtains an accuracy of 84.25%, this indicates that by using genetic algorithms for the selection of features, it is possible to select the representative ones without affecting the performance of the classifier, which in this case was Naïve Bayes.

3 Method and Materials

According to the literature review, in the present investigation, a genetic algorithm was developed with a support vector machine as an aptitude function that allows selecting the most significant features using the lexical features (model of n-grams at character level) applied to the authorship attribution task to improve the classification. As observed in the state of the art, the learning method that obtains the best results in the classification is the support vector machine [11, 29].

The feature selection method that has been implemented in particular for the C10 corpus is the most frequent term filter method. This type of method selects the features according to a relevance measurement, this selection is carried out empirically. However, the wrapper methods do not perform the selection empirically, they consider the entire possible set of solutions. The best subsets of features are selected from supervised learning.

The proposed method for the selection of features consists of the development of three stages.

1. Development of a text classification method.
2. Genetic algorithm for the selection of features.

Table 1. Genetic algorithm parameters.

Poblation size	Selection	Cross	Bitwise mutation	Elitism
100 individuals	Deterministic Tournament	Uniform at 80%	0.01%	3

3. Evaluation of the features obtained with the genetic algorithm.

Text classification method: the method to be implemented consists of the following stages [31]:

- Data acquisition: the data set used in this research is available on the PAN website (https://pan.webis.de/) this corpus is used for the detection of plagiarism and currently also the authorship identification. The C10 corpus is used for the task of authorship attribution in different works of literature.
- Data analysis and labeling: In this stage, the features extraction process is carried out for each author's document. Later they are represented by the vector model.
- Construction of features and weighting: Boolean or binary weighting assigns a value of one when the term is present and a zero in the absence of the value, assigning these values in the vector representation [31, 32].
- Selection and projection of features: In the implementation with the genetic algorithm no method of selection of features is used, and for the evaluation of the features those obtained by the genetic algorithm are used.
- Training of a classification model: The learning method used is a support vector machine using a linear kernel, the parameter C equal to one. The support vector machine in the genetic algorithm is trained with the features of each of the authors, generating a model that allows an evaluation to be carried out.
- Evaluation of the solution: The metric used is accuracy, it is defined in terms of true positives (TP), False positives (FP), True negatives (TN), and False negatives (FN) as shown below [29].
- $\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$.

3.1 Selection of Features with the Genetic Algorithm

For the selection of features with the genetic algorithm, only the training documents are used, these are divided into a new set of training and validation, to be able to carry out the evaluation with the support vector machine. The representation used for individuals is a binary encoding, where when the value of the gene is one the feature is considered, on the other hand, if the value is zero it is discarded.

Rewriting the documents with the features to be used so that the individual can be evaluated with the support vector machine using the accuracy metric to know the individual's aptitude.

The operators used by the genetic algorithm are described below, these were determined according to the analysis presented in the experimentation section:

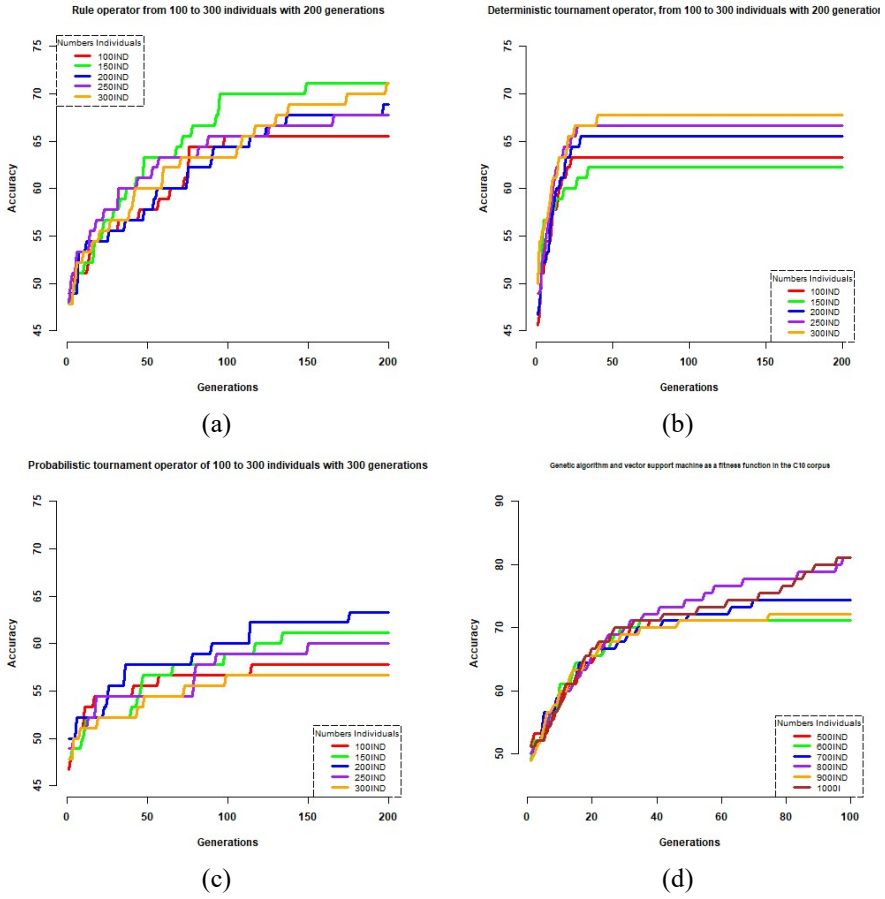

Fig. 1. Results of genetic operators.

3.2 Evaluation of the Features

The evaluation was carried out on the set of features considered the most significant by the genetic algorithm, starting from the fittest individual. The process that is carried out for the evaluation is as follows:

1. The training and validation documents are featured with the 3-gram model.
2. Boolean weighting and vector representation are performed with the features selected by the genetic algorithm.
3. The support vector machine training is performed.
4. From the model generated with the support vector machine, the validation documents are evaluated with the accuracy metric.

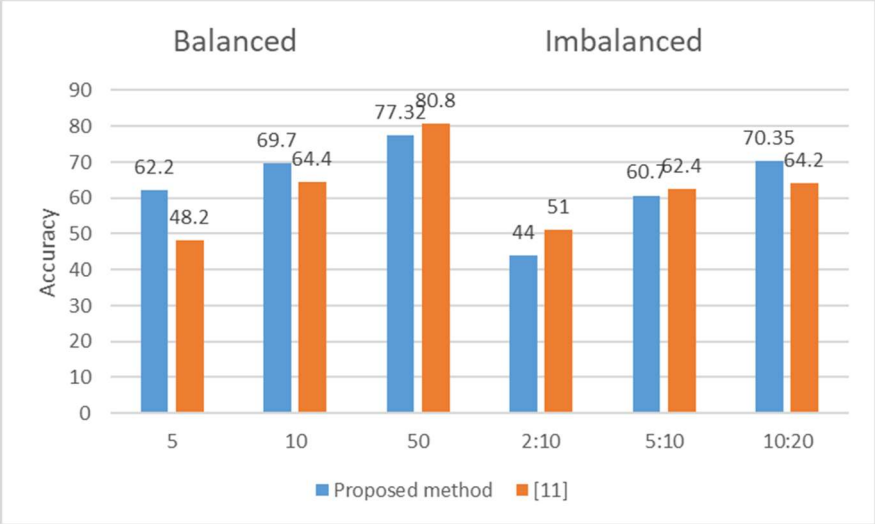

Fig. 2. Analysis of Balanced and Imbalanced samples.

4 Experimentation

The corpus used in the experimentation is the C10 corpus. An analysis of the different operators is required for a genetic algorithm with binary coding since the genetic algorithm implemented in the present investigation has an average of 5,000 genes as explained in the method each individual is a text classification process, which generates a high computational cost. Fig. 1 shows an analysis to determine the best operators for the genetic algorithm.

For the experimentation, the following parameters of the genetic algorithm were considered. Genetic operators; selection operator: roulette, deterministic and probabilistic tournament, crossover operator: uniform with 80%, the mutation operator has a probability of 0.01%. The experimentation to determine the appropriate parameters was carried out with a sample of size 10 from Corpus C10.

a. Analysis with the roulette operator, from 100 to 300 individuals with 200 generations.

For the first experiment, the following population sizes were considered; 100, 150, 200, 250, and 300 individuals (IND). As can be seen in Fig. 1a, the evolution process is slower in larger populations than in smaller populations, this is due to the fact that in smaller populations the individuals with better aptitude tend to be selected more frequently, but it is also observed in Fig. 1a shows that the genetic algorithm with this selection operator requires a greater number of generations.

b. Analysis with the deterministic tournament operator, from 100 to 300 individuals with 200 generations.

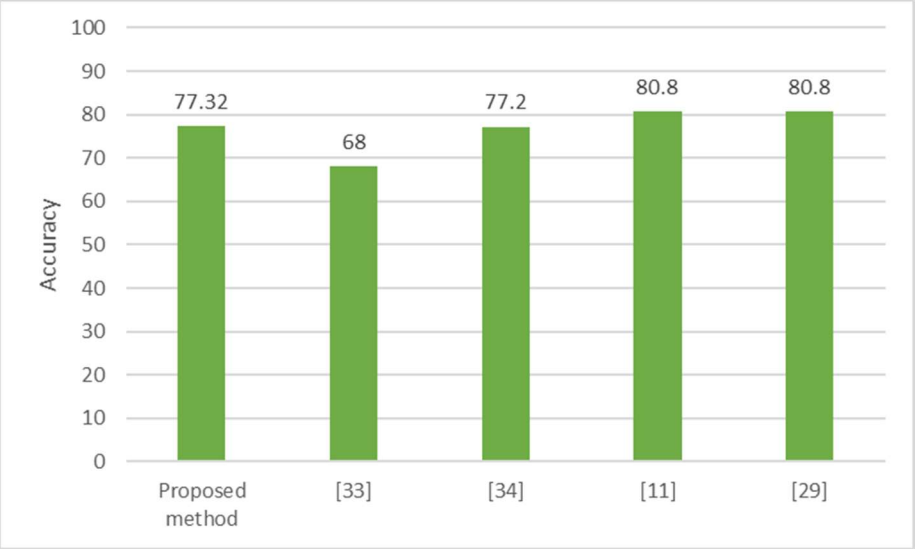

Fig. 3. Proposed method versus the state of the art.

For the second experiment, the following population sizes were considered; 100, 150, 200, 250, and 300 individuals (IND). As can be seen in Fig. 1b, the process of evolution is faster, requires fewer generations, and the greater the number of individuals in the population, the better results are obtained, because in larger populations the individuals with better aptitude tend

To be selected more frequently, according to the experimentation carried out, it is better to have more individuals than to have more generations with this operator, the computational cost is lower when the number of individuals is increased compared to the increase in generations.

c. Analysis with the probabilistic tournament operator, from 100 to 300 individuals with 200 generations. For the third experiment, the following population sizes were considered; 100, 150, 200, 250, and 300 individuals (IND). As can be seen in Fig. 1c, the smaller the

population, the better the evolution. With older populations, evolution is gradual, but it requires a greater number of generations and a greater number of individuals, which would imply a high computational cost.

d. Deterministic tournament operator analysis, from 500 to 1000 individuals with 100 generations.

Based on the previous experimentation, it was determined that the selection operator for the present investigation is a deterministic tournament, it presents a better behavior, and it requires a smaller number of generations.

As shown in Fig. 1d, the more individuals the population has, the better the evolution, for this reason, it was determined to use 100 generations with 1000 individuals in the experiments.

e. Selection of features with a genetic algorithm and vector support machine as a fitness function in the C10 corpus.

The results obtained from the present investigation are compared with the results of [11] for balanced and imbalanced samples.

As can be seen in Fig. 2, the best results were obtained in two out of three samples, this indicates that the genetic algorithm with a support vector machine as an aptitude function is a good method for the selection of features; in the 5 sample the accuracy is improved by 14.0%, and in the 10 sample the accuracy is improved by 5.3%.

However, in the 50 sample, the proposed method is exceeded by 3.48%. As can be seen in Fig. 2, better results were obtained in one of the samples, in the 10:20 sample, the accuracy is improved by 6.15%.

However, in samples 2:10 and 5:10, the proposed method is exceeded by 7% and 1.7% respectively; the method with imbalanced samples requires that the training sample be larger in order to improve the performance of the classifier.

As can be seen in Fig. 3, the proposed method is compared with other methods used for authorship attribution, such as the proposal in [33] that performs the construction of syntactic graphs, in [29] uses a representation of n-grams of words with Doc2vec, in [34] implements the 15 most referenced works about the authorship attribution task to replicate the methods and analyze whether the methods are replicable.

5 Conclusions

The selection of features does impact the performance of the classifier as shown in Fig. 2a and 2b. In balanced samples, better results are obtained in two of the three samples, improving the accuracy by 5.3% in the 10:10 sample. and 14.0% in the 5:5 sample.

In the imbalanced samples, better results are obtained in one of the three samples, improving the accuracy by 5.65% in the 10:20 sample, when changing the method of selection of features, better results are obtained in three of the six established samples.

The proposed method selects the features by means of a learning method.

According to the experimentation carried out in the present investigation, the following conclusions of the implemented method are determined:

– A genetic algorithm and support vector machine as a fitness function allows selecting a relevant set of features.

– Table two shows the results obtained with the validation documents where it is observed that the method proposed in this research obtains competitive results with the state of the art.

The main contribution of this work is the development of a feature extraction method for the authorship attribution task with a genetic algorithm that implements support vector machine as an aptitude function, this method can be implemented in different language processing tasks natural to reduce dimensionality, improve classification, and identify which are the most significant features.

From the present investigation, an area of opportunity arises with the implementation of a micro heuristic that allows reducing the computational cost of the genetic algorithm as a method for the selection of features.

References

1. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, vol. 60, no. 3, pp. 538–556 (2009) doi:10.1002/asi.21001
2. Lambers, M., Cor, J.: Forensic authorship attribution using compression distances to prototypes. Computational Forensics, Springer Berlin Heidelberg, pp. 13–24 (2009) doi:10.1007/978-3-642-03521-0_2
3. Pillay, S. R., Thamar, S.: Authorship attribution of web forum posts. ECrime Researchers Summit (2010) doi:10.1109/ecrime.2010.5706693
4. Ramnial, H., Panchoo, S., Pudaruth, S.: Authorship attribution using stylometry and machine learning techniques. Advances in Intelligent Systems and Computing, pp. 113–125 (2015) doi: 10.1007/978-3-319-23036-8_10
5. Zhang, Z., Li, X., Tian, X.: Research on feature weights of Liheci word sense disambiguation. In: Proceedings 8th International Symposium on Computational Intelligence and Design, vol. 2, pp. 7–10 (2015) doi: 10.1109/ISCID.2015.221
6. Shrestha, P., Sierra, S., González, F. A., Montes, M., Rosso, P., Solorio, T.: Convolutional neural networks for authorship attribution of short texts. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, pp. 669–674 (2017)
7. Kern, R., Seifert, C., Zechner, M., Granitzer, M.: Vote/veto meta classifier for authorship identification. In: Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (2011)
8. Pillay, S. R., Solorio, T.: Authorship attribution of web forum posts. In: 2010 eCrime Researchers Summit, pp. 1–7 (2010) doi: 10.1109/ecrime.2010.5706693
9. Anwar, W., Bajwa, I., Ramzan, S.: Design and implementation of a machine learning based authorship identification model. Scientific Programming, vol. 2019, pp. 1–14 (2019) doi: 10.1155/2019/9431073
10. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. In: Association for Computing Machinery Special Interest Group on Management of Data, vol. 30, no. 4, pp. 55–64 (2001) doi: 10.1145/604264.604272
11. Plakias, S., Stamatatos, E.: Tensor space models for authorship attribution. In: Darzentas, J., Vouros, G.A., Vosinakis, S., Arnellos, A. (eds) Artificial Intelligence: Theories, Models and Applications. SETN 2008. Lecture Notes in Computer Science, Springer, Heidelberg, vol. 5138, pp. 239–249 (2008)
12. Xue, B., Zhang, M., Browne, W.: A comprehensive comparison on evolutionary feature selection approaches to classification. International Journal of Computational Intelligence and Applications, vol. 14, no. 2 (2015) doi: 10.1142/S 146902681550008X
13. Tan, F., Fu, X., Zhang, Y., Bourgeois, A.: A genetic algorithm-based method for feature subset selection. Soft Computing a Fusion of Foundations, Methodologies and Applications, vol. 12, no. 2, pp.111–120 (2008) doi: 10.1007/s00500-007-0193-8
14. Mohamad, M., Deris, S., Mat, S., Razib, M.: Feature selection method using genetic algorithm for the classification of small and high dimension data. In: Proceeding of the First International Symposium on Information and Communications Technologies, pp. 13–16 (2004)
15. Mikros, G., Perifanos, K.: Authorship identification in large email collections: Experiments using features that belong to different linguistic levels. Notebook for PAN at CLEF (2011)
16. Vilariño, D., Castillo, E., Pinto, D., León, S., Tovar, M.: Baseline Approaches for the Authorship Identification Task. Notebook for PAN at CLEF 201. (2011)
17. Kern, R., Seifert, C., Zechner, M., Granitzer, M.: Vote/veto meta classifier for authorship identification. In: Proceedings of the 2011 Conference on Multilingual and Multimodal

Information CLEF´11: Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands (2011)
18. Vartapetiance, A., Gillam, L.: Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. CLEF 2012 Evaluation Labs and Workshop Working Notes Papers (2012)
19. Mesleh, A.: Chi square feature extraction based SVMS Arabic language text categorization system. Journal of Computer Science, vol. 3, no. 6, pp. 430–435 (2007)
20. D'Andrea, A., Ferri, F., Grifoni, P., Guzzo, T.: Approaches, Tools and Applications for Sentiment Analysis Implementation. International Journal of Computer Applications, vol. 125, no. 3 (2015)
21. Varela, P., Martins, A., Aguiar, P., y Figueiredo, M.: An empirical study of feature selection for sentiment analysis. In: 9th conference on telecommunications (2013)
22. Adel, A., Omar, N., Al-Shabi, A.: A comparative study of combined feature selection methods for Arabic text classification. Journal of Computer Science, vol. 10, no. 11 (2014) doi: 10.3844/jcssp.2014.2232.2239
23. Pavlyshenko, B.: Genetic optimization of keywords subset in the classification analysis of texts authorship. In: Journal of Quantitative Linguistics, vol. 21, pp. 341–349 (2014) doi: 10.48550/ARXIV.1211.3402
24. Batista, B., Moreno, J., Moreno, M.: Algoritmos genéticos, una visión práctica, números. Revista de Didáctica de las Matemáticas, vol. 71, pp. 29–47 (2009)
25. Coello, C.: Introducción a los algoritmos genéticos. Soluciones Avanzadas. Tecnologías de Información y Estrategias de Negocios, vol. 3, no. 7, pp. 5–11 (1995)
26. Gestal, M., Rivero, D., Rabuñal, J., Dorado, J., Pazos, A.: Introducción a los algoritmos genéticos y la programación genética (2010)
27. Ponce, P.: Inteligencia artificial con aplicaciones a la ingeniería (2010)
28. Su, S., Li, L., Zhao, Q.: Text feature selection based on improved adaptive GA. In: 7th International Conference on Natural Language Processing and Knowledge Engineering, pp. 169–172 (2011)
29. Posadas-Durán, J., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., Chanona-Hernández, L.: Application of the distributed document representation in the authorship attribution task for small corpora. Soft Computing, vol. 21, pp. 627–639 (2017)
30. Mirończuk, M., Protasiewicz, J.: A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, vol. 106, pp. 36–54 (2018) doi: 10.1016/j.eswa.2018.03.058
31. Ledeneva, Y., García, R.: Generación automática de resúmenes: retos, propuestas y experimentos (2017)
32. Zhang, Z., Li, X., Tian, X.: Research on feature weights of Liheci word sense disambiguation. In: Proceedings of 8th International Symposium on Computational Intelligence and Design, vol. 2, pp. 7–10 (2015) doi: 10.1109/ISCID.2015.221
33. Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., Gelbukh, A.: Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. Sensors, vol. 16, no. 9 (2016) doi: 10.3390/s16091374
34. Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J.M., Köhler, J., Lötzsch, W., Müller, F., Müller, M.E., PaBmann, R., Reinke, B., Retenmeier, L., Rometsch, T., Sommer, T., Träger, M., Wilhem, S., Stein, B., Stamatatos, E., Hagen, M.: Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. In: Proceedings of European Conference on Information Retrieval. Lecture Notes in Computer Science, vol. 9626, pp. 393-407 (2016) doi: 10.1007/978-3-319-30671-1_29