View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Servicio de Difusión de la Creación Intelectual

An experimental study for the Cross Domain Author Profiling classification

María José Garciarena Ucelay, María Paula Villegas, Leticia Cecilia Cagnina, Marcelo Luis Errecalde

Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Facultad de Ciencias Físico, Matemáticas y Naturales,
Universidad Nacional de San Luis – Ejército de los Andes 950
(D5700HHW) – San Luis – Argentina, Tel.: (0266) 4420823 / Fax: (0266) 4430224
emails: {mjgarciarenaucelay, villegasmariapaula74, lcagnina, merrecalde}@gmail.com

Abstract. *Author Profiling* is the task of predicting characteristics of the author of a text, such as age, gender, personality, native language, etc. This is a task of growing importance due to the potential applications in security, crime detection and marketing, among others. An interesting point is to study the robustness of a classifier when it is trained with a dataset and tested with others containing different characteristics. Commonly this is called *cross domain* experimentation. Although different cross domain studies have been done for datasets in English language, for Spanish it has recently begun. In this context, this work presents a study of cross domain classification for the author profiling task in Spanish. The experimental results showed that using corpora with different levels of formality we can obtain robust classifiers for the author profiling task in Spanish language.

Keywords: Author Profiling, Natural Processing Language, Cross Domain Classification

1 Introduction

The evolution of the World Wide Web sites to the Web 2.0 has mainly implied a proliferation of contents created and shared from all kinds of users in different social networks. Also, it has facilitated the increment of falsification of identity, plagiarism and a significant increase in the traffic of spam data through the Internet. For this reason, automatic methods are needed to detect if a given text belongs to a specific author, if the gender and age stated by a user of social media is compatible with his/her writing style, etc. In this context, the Author Profiling task refers to the identification of different demographic aspects like gender [1], age [2, 3], native language [4], emotional state [5, 6] or personality [5, 7] of an anonymous author of a text [8].

A particular problem concerned with the author profiling task in Spanish language is the lack of data for experimentation. For that, it is important to take in advantage of all the available data in order to obtain good and enough general classifiers for the task and then, to use those for new data that can be collected.

Traditional machine learning methods construct reliable and accurate models using available labeled data. These models are generally tested with data drawn from the underlying distribution or domain. Then, a classification model working well in one domain could not work as well in another one [9]. *Cross domain* classification is used to tackle that problem.

For *domain* we can consider the source of the documents (Twitter, blogs, chats, magazines, news, etc) [9], topics (places, politic, food) [10], products (books, furniture, movies) [11], research areas (computer science, biology, physics) [11], etc. In the present work we define the domain such as the level of "informality" of a text.

In PAN-2014 competition an extra experiment of cross domain was held, for both English and Spanish languages, which served as a previous work [12]. Thus, here we perform several experiments in order to determine the corpus we can obtain a general classifier with.

In this paper we present the results obtained from carrying out cross domain experiments. Such tests have not been previously performed in Spanish due to the lack of resources in this language and because training with a corpus and then testing with another is a recently studied approach. However, cross domain experimentation becomes an interesting field for researchers working in actual classification tasks as author profiling is. We have used available corpora provided for PAN competitions (2013 and 2014) which present a high level of informality in the texts contained. Also we have considered a formal corpus named SpanText [13] with similar characteristics with respect to those of PAN competitions, in terms of genre and age of people who wrote the texts. The results obtained with the experimentation with cross domain in terms of informality of the texts demonstrate that reliable classifiers can be obtained for the author profiling classification task.

The cross domain experiments may be helpful for other tasks, besides contributing to the author profiling itself. For example, it could be used to generate a classifier from a large collection of different types of texts, properly selected. Then, that classifier could be used to analyze texts hard to obtain or for analyzing online data.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the author profiling task and some concepts related to cross domain experiments. In Section 3, the main characteristics of the different data collections used in the experimentation are presented. Section 4 describes an experimental study about cross domain among the available corpora in Spanish language. Finally, in Section 5 some conclusions are drawn and future works are proposed.

2 Author Profiling Task

Nowadays, the evolution of the Web sites on the Internet and the increasing use of social networks like Facebook and Twitter have made available a huge amount of information. A large part of this information is in plain text and it can be used to infer about the writer. The Author Profiling Task (APT) consists in knowing as much as possible about an unknown author, just by analyzing the given text [5]. In this regard, profiling tries to determine the author's gender, age, level of education, geographic origin, native language and personality type [1-8].

The APT has mainly focused on documents written in English but, according to our knowledge, this situation has started to change with papers presented at PAN-2013 competition [14], when the organizers considered the gender and age aspects of the author profiling problem, both in English and Spanish.

Now, we have several collections in Spanish with different kind of "formality". That is, a corpus is "informal" if the texts have noise like typos, images, hyper-links, emoticons, contractions, etc. This noise becomes the corpus in a very challenging dataset for any classifier. However, from the results of the competition PAN-2013 it can be seen that some approaches like the one used in [15] (the winner of the competition), can obtain interesting results even when the nature of the documents makes very difficult the classification. Unfortunately, it is unclear how these techniques work when these are trained with some corpora and tested with data with a different distribution. This was the reason that motivated us to study the cross domain approach.

When we talk about a *domain*, in the data mining field, it could be loosely defined as a specialized area of interest for which we can develop ontologies, dictionaries and taxonomies of information. We can refer to the different scopes (very broad or more narrowly specialized domains), or also the type of source from which the texts come from (like blogs, forums, etc.), or simply, a domain could be considered as the writing style (formal, informal, scientific, etc.). Thus, cross domain can be interpreted in several ways.

However, this paper simply assumed that a *domain* is a texts collection with a particular level of informality. Therefore, a *cross domain* experiment indicates a classification where you train with a corpus with certain level of informality and test with other with a different level of informality. Cross domain tests are also called by others authors as Domain Transfer experiments [16]. These consist in generating a classifier from texts that belongs to a source domain (training set) to apply it to a different target domain (test set). In other words, the underlying purpose of this concept is to check how well the trained classifier generalizes when it run on a different collection of documents.

3 Data Collections

We consider three different corpora in Spanish for the experimental study: *SpanText* and others two which were provided by the PAN-CLEF competition in the years 2013 [14] and 2014 [12]. These latter are called *PAN-2013* corpus and *PAN-2014* corpus. Also, we use a sub-corpus of PAN-2013 which we have proposed for this experimentation. The characteristics of each one are presented below.

SpanText is a set of "formal" documents written in Spanish extracted from the Web [13]. In this context, we use the term "formal" (as opposed to "informal") to refer to those documents whose content has a low percentage of "non-dictionary" words, abbreviations, contractions, emoticons, slang expressions, etc. that are typical in messaging and the social Web. This dataset consists of a variety of texts that one supposes to find in newspapers, students' reports, books and so on. These "speak" about different topics and they were written by Spanish speakers from Spain and Latin American countries. Besides, there are only one document (file) per author.

Two versions of this collection were presented in [13]. They are called "balanced" and "unbalanced" versions. However, there is another one called "semi-balanced", in which we are interested. Spantext (like PAN-2013) considers age and gender as the basic demographic information for the authors. All the documents are labeled with both characteristics. For age detection, it contemplates three classes: 10s, 20s and 30s. In the semi-balanced version, the number of documents per class is proportional to the amount of PAN-2013's documents. These are only uniformly distributed with respect to gender.

Regarding the PAN-2013 collection, it was built automatically with texts from blogs and other social networks [14]. The organizers of the competition provided two corpora: one in English and other in Spanish language. The dataset was divided into the following sub-sets: training, early bird evaluation and final testing. In this work, PAN-2013 will refer to the training and test sets of the Spanish language. Documents in PAN-2013 considered a wide spectrum of topics and they include "informal" text. The posts were grouped by author selecting those authors with at least one post and chunking in different files with more than 1000 words in their posts. But it also included some authors with few and shorter posts. For age classification, this collection considers the three same classes as SpanText and it is balanced by gender and imbalanced by age group, having more texts in class 20s than in 30s, and more in 30s than in 10s.

However, due to the difference between the sizes of SpanText and PAN-2013, it was needed to separate a sub-corpus of the latter (called sub-PAN2013), so it has the same number of documents per category as the semi-balanced version of the former. Thus, the results of diverse experiments can be fairly compared and the difference in the results will be limited to other variables, such as the quality of the texts.

The collection of texts written in Spanish in the PAN-2014 corpus was collected semi-automatically from four different sources: social media, blogs, Twitter and hotel reviews (the last only provided in the English corpus). In the competition of the year 2014, the PAN-CLEF organization opted for modeling age in a more fine-grained way and considered the following ranges (classes): 18-24, 25-34, 35-49, 50-64 and 65+ years old. The full collection was also divided into training, early bird evaluation and final testing parts. It is worth noting that we could access only to the training set and we use that part in the experimental study because the test corpus is not available at the time of writing this article.

Table 1. Vocabulary (number of words without repetition), number of terms (words), number of files and average number of terms for each collection.

Collection	SpanText	PAN-2013	PAN-2013 sub-corpus	PAN-2014
#Vocabulary	31 504	342 068	29 616	306 809
#Terms	294 434	22 868 586	294 596	17 686 634
#Files	1 000	84 060	1 000	1 500
Average Terms	294	301	294	11 806

Table 1 shows the statistics for each full collection. We can observe that PAN-2013 corpus presents the biggest numbers except in average number of words. This is because of its structure, it has more files (or authors) and more wealth in terms of writing styles but, the texts are not too long.

If we compare SpanText with PAN-2014, the latter is 50% bigger than the former, but SpanText only has a 10% of vocabulary than its counterpart. PAN-2014 prioritized the amount of texts from the same author, rather than the number of authors. It was probably because these are often short texts due to the source from which they came from (e.g. Twitter). This is verified in the amount of average terms for document that overcomes highly the other two corpora.

However, we must emphasize that in this regard SpanText is not far from the PAN-2013 collection. Perhaps if we could increase the number of documents of SpanText, maintaining its characteristics, this corpus would become the most useful. Since the proportion between, the amount of repeated words and the vocabulary is 10% for SpanText and 1% for both PAN-2013 and PAN-2014.

4 Experimental Study

In this section, we describe the cross domain experiments performed using the software WEKA [17]. Basically we performed two kinds of studies: APT as a classification of documents by gender, and then, considering both together age and gender. This is because, as we previously mentioned, the corpus PAN-2014 considered different age ranges from the ones defined in PAN-2013; in such way that we cannot make a join or separation of categories in order to consider the same ranges of age for both corpora.

Table 2(a) shows the information about the cross domain experiments: name of the corpus used for training and amount of documents considered, and name of the corpus used for testing with the corresponding amount of documents for performing the classification only by gender. The same information for the classification by gender and age considered together is shown in Table 2(b). From now on, to refer to a particular experiment, first we will mention the name of the corpus that was used to train, followed by the name of the collection employed to test (short forms of the original names of the corpora). For example, SPAN-PAN13 corresponds to the experiment which uses SpanText to generate the model and PAN-2013 to validate it.

Table 2. List of the cross domain experiments carried out.

(a) Classifications only by gender

Training	Docs	Test	Docs
PAN-2014	1 500	SpanText	1 000
SpanText	1 000	PAN-2014	1 500
PAN-2014	1 500	PAN-2013	84 060
PAN-2013	84 060	PAN-2014	1 500
SpanText	1 000	PAN-2013	84 060
PAN-2013	84 060	SpanText	1 000

(b) Classifications by age and gender

Training	Docs	Test	Docs
PAN-2013	84 060	SpanText	1 000
SpanText	1 000	PAN-2013	84 060
Sub-PAN13	1 000	SpanText	1 000
SpanText	1 000	Sub-PAN13	1 000

We used two traditional models of representation of documents: *bag of words* (BoW) [18] and *character trigrams* [19]. Regarding the weighting schema, we employed: *Boolean* [18] and *tf-idf* [20]. We also considered the *Second Order Attributes* (SOA) representation [13] because it has been demonstrated to be effective for this task. We have constructed the models and performed the classification using N*aïve Bayes* [21] and *LibLINEAR* [22] methods. Besides those, we considered an interesting approach *Sistema de Perfiles* (SP) [23] which generates its own model (profiles) using the most frequent character trigrams of the texts (L value) and then evaluates the belonging of the test documents in the profiles. It is important to note that due to the characteristics of its functioning, we could not use SP for those experiments which required to train with the PAN-2014 collection, because it was not able to generate the required profiles for the classification. The values for the L parameter of SP mentioned in the tables were chosen from carrying out prior executions for different values of this, choosing the one with we obtained the best accuracy. All approaches were evaluated considering the accuracy as metric.

4.1 Classifications only by gender

The percentages of correctly classified instances (accuracy) obtained in the cross domain classification only by gender are shown in Table 3. The table is divided into three sub-tables (a), (b) and (c) considering three different cross domain experiments. The highest accuracy values obtained are highlighted in boldface. The first value is the accuracy obtained with Naïve Bayes algorithm and the one after the slash corresponds to the accuracy obtained with the LibLINEAR algorithm.

Table 3. Accuracy obtained in cross domain classifications only by gender with "Naïve Bayes / LibLINEAR" algorithms.

	(a) PAN-2014 and SpanText		(b) PAN-2013 and SpanText		(c) PAN-2014 and PAN-2013	
	PAN14-SPAN	SPAN-PAN14	PAN13-SPAN	SPAN-PAN13	PAN13-SPAN	SPAN-PAN13
Boolean words	50,0 / 48,2	54,1 / 52,6	53,0 / 58,1	**53,1** / 52,1	50,3 / 52,4	57,5 / **67,5**
TF-IDF words	51,7 / 53,6	52,7 / 52,9	51,6 / **60,9**	51,1 / 51,9	52,4 / 53,3	58,3 / 64,9
SOA words	**61,2** / 59,4	54,9 / **55,4**	60,1 / 53,0	50,1 / 50,0	**59,1** / 58,5	61,1 / 62,6
Boolean 3grams	50,0 / 49,8	48,9 / 51,5	51,0 / 55,8	**57,7** / 51,6	50,1 / 49,5	50,7 / 58,8
TF-IDF 3grams	50,1 / **53,7**	51,3 / 51,3	54,8 / **60,3**	50,6 / 50,1	**54,9** / 54,6	56,3 / **62,2**
SP 3grams	-	**54,3**	58,7	51,3	-	57,8

The baseline used by PAN-CLEF Lab competition to determine if a two-class classifier is acceptable is 50%. Table 3 shows that almost all percentages exceeded or

equaled this value (48,2; 49,8; 48,9 and 49,5 are the exception). Note that with PAN13-SPAN it was not obtained percentages lower than the 50%.

Figure 1 provides a visual summary of Table 3. The bars with no plot at the left correspond to the representation of documents and the bars with plot (dots and rhombus) at the right with classifiers. The accuracy shown is the average of all the accuracies obtained for each approach for every training corpus used. Furthermore, results are shown from the baseline so that it would highlight better the differences obtained. It is important to note that words strategies dominate character trigrams approaches.

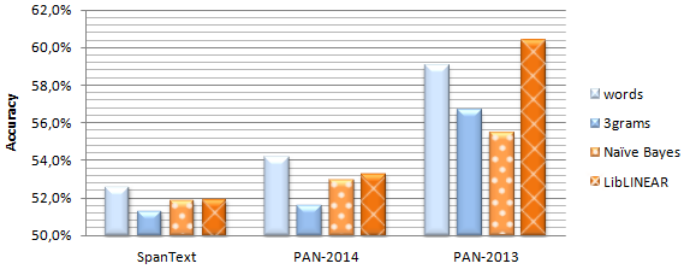
If we analyze the document representations, in general SOA accomplish the best performing, which precisely works with words. Next it follows the SP with character trigrams, then, in third and fourth places are the tf-idf representation with words and character trigrams respectively. Certainly, with a little more elaborated approaches than the simple use of frequencies, it achieves better results.

Regarding the classifiers, it can be concluded that LibLINEAR is superior. Out of the eleven best results in bold, six were obtained with it. Moreover, if an average of executions is calculated grouping them by classifier, it results that using LibLINEAR the average accuracy is around 60%, while with Naïve Bayes reaches only 56%.

The highest results were achieved when we trained with PAN-2013 and tested with PAN-2014, 67.5% for words and 62.2% for character trigrams. If we make an average of all executions in which this corpus was used to train the model, we found that this obtained the best percentage. This is also exhibited in Figure 1. Therefore, with 57.8% against 52.9% training with PAN-2014 and 51.9% with SpanText, we can say that the PAN-2013 collection is the one that generates a more general classifier.

At the PAN-CLEF competition in 2014, they tested the approaches of the participants who participated in 2013 (the approaches were trained with PAN-2013 corpus) using the 2014 collection (testing with PAN-2014). The SP achieved 69.4% of accuracy taking the first position in the final ranking [12]. Observing the results obtained we conclude that with the PAN-2013 collection we can get a general model able to classify documents from different corpora. Additionally, the results of the experiments accomplished in this work, at least for classifications only by gender are promising and overcome at least in a 3% the experiments performed on a single domain.

Figure 1. Summary of the results obtained for the cross domain classification only by gender distinguished by representations and classifiers.



4.2 Joint classifications by age and gender

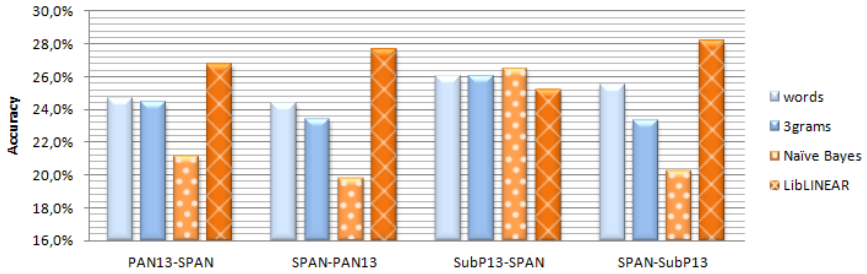
The results obtained for the cross domain classification considering age and gender are shown in Table 4. The best result of each section is highlighted in boldface. The baseline for this case is 16% because there are six categories (the combination of female and male with the three ranges of age). In Table 4 there are three cases in which the percentage does not reach the baseline. The first correspond to SPAN-PAN13 combination employing words-tf-idf representation and Naïve Bayes. Then, the second and third cases use character trigrams-Boolean representation with Naïve Bayes classifier, SpanText to train and PAN-2013 (or its sub-corpus) to test. However, when an average of the results is calculated, for example based on the classifiers, we can say that all the values are over the baseline.

Table 4. Accuracy obtained in cross domain classifications by age and gender with "Naïve Bayes / LibLINEAR" algorithms.

	(a) SpanText and PAN-2013		(b) SpanText and sub-PAN2013	
	PAN13-SPAN	SPAN-PAN13	subP13-SPAN	SPAN-subP13
Boolean words	19,7 / 26,5	20,1 / 28,1	20,0 / 25,7	19,8 / 28,3
TF-IDF words	19,3 / **29,5**	14,0 / 28,3	24,0 / 24,8	17,5 / **29,6**
SOA words	25,6 / 27,4	**28,7** / 27,1	**35,1** / 26,4	29,2 / 28,8
Boolean 3grams	20,5 / 22,8	11,5 / **28,0**	25,0 / 23,7	15,1 / 26,9
TF-IDF 3grams	20,7 / 27,7	24,8 / 26,9	**28,4** / 25,5	19,8 / 27,3
SP 3grams	**30,5**	25,9	27,5	**27,6**

Figure 2 summarizes the information of Table 4. In the bars the different models of representation (words and character trigrams) are at the left and they do not have a plot. Whereas the bars with dots and rhombus that are at the right, represent the behavior of the classifiers. As compared to the cross domain classifications only by gender, where words always predominated, here the bars exhibited are more similar among them. So it seems that the character trigrams help to distinguish better out the six categories.

Figure 2. Summary of the results obtained for cross domain classifications by age and gender distinguished by representations and classifiers.



If we analyze the traditional representations, i.e. Boolean and tf-idf, we obtained better results using words when we trained with the complete PAN-2013 corpus (Table 4 (a)). In particular, the combination of the tf-idf representation with the classifier LibLINEAR has worked considerably well. However, when it is trained with SpanText, the character trigrams strategy achieves a higher percentage on average. Now if we consider slightly more elaborated approaches in SPAN-PAN13 combination, the SOA representation is the best at discriminating the different classes. Nevertheless, the best overall result for the joint classification by gender and age is reached in PAN13-SPAN with the SP.

Table 4 (b) shows the results obtained with the sub-corpus of PAN-2013 which are different than those obtained with the complete corpus of PAN-2013. Even though, this case is a specific one thereof.

In general, regarding the classifiers, Naïve Bayes obtained poor results, highlighting even more the difference in performance respect to its counterpart. As we mentioned above, LibLINEAR with tf-idf representation using words obtained the second best result for cross domain classification by gender and age using the whole corpora.

Thus, in these experiments the same behavior is observed as in the classifications only by gender in which the approaches that use words are better. This is evidenced by the 35.1% obtained with the SOA representation in PAN13-SPAN combination. In addition, the highest percentage is accomplished again using the sub-corpus PAN-2013 to train the model.

5 Conclusions and Future Work

Cross domain experimentation has started to raise the interest of researchers turning their attention to the possibility of building a general enough classifier to classify any type of text documents. Hence its importance in the APT in which it is difficult to find properly labeled and lesser noise collections of texts, particularly for the Spanish language, is significant. For example, to detect pedophiles on the network or other kind of tasks that require a real-time response, and where the previous training with information which is not necessarily of the same type of the task to evaluate, is limited or non-existent.

In this paper we present a preliminary study considering cross domain author profiling classification. We made different experiments considering some corpora for training and testing using others considering different level of formality.

We analyzed the corpora available for APT in Spanish language using different representations and classification algorithms. Aiming not only to see how well a corpus generalizes a model, but also to evaluate the desirable characteristics that should have them, we conclude that the PAN-2013 collection is the one which better serves for that purpose. The highest accuracies were obtained with more elaborate representations such as SOA and approaches such as SP. Therefore, the results of the cross domain experiments obtained in this study turn to be promising, since they get close and even exceed the values obtained in experiments conducted in a single domain (or inter-domain).

Finally, it would be interesting to verify how the SP approach would behave when it trained with the PAN-2014 collection, and instead of using character trigrams, using words or more sophisticated representations.

References

1. Koppel, M., Argamon, S., and Shimoni, A. R. Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing. Vol. 17, no 4, pp. 401–412, 2002.
2. Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. Gender, Genre, and Writing Style in Formal Written Texts. TEXT. Vol. 23, pp. 321–346, 2003.
3. Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. Effects of Age and Gender on Blogging. In AAAI Spring Symposium: Comp. Approaches to Analyzing Weblogs. Vol. 6, pp. 199–205, 2006.
4. Koppel, M., Schler, J., and Zigdon, K. Determining an Author's Native Language by Mining a Text for Errors. In Proc. of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. ACM, pp. 624–628, 2005.
5. Rangel, F. Author Profile in Social Media: Identifying Information about Gender, Age, Emotions and beyond. Proc of the 5th BCS IRSG Symposium on Future Directions in Information Access, pp.58–60, 2013.
6. Bo, P., and Lee, L. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. Vol. 2, issue 1-2, pp. 1–135, 2008.
7. Pennebaker, J.W., Mehl, M.R., and Niederhoffer, K. Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology. Vol. 54, pp. 547–577, 2003.
8. Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. Automatically Profiling the Author of An Anonymous Text. Communications of the ACM. Vol. 52, pp. 119–123, 2009.
9. Ramakrishna Murty, M., Murthy, J.V.R., Prasad Reddy, P.V.G.D., and Satapathy, S.C. A survey of cross-domain text categorization techniques. In Recent Advances in Information Technology (RAIT), 1st International Conference, pp. 499–504, 2012.
10. Li, L., Jin, X., and Long, M. Topic Correlation Analysis for Cross-Domain Text Classification. In Proc. AAAI, 2012.
11. Pan, S. J., Ni, X., Sun, J., Yang, Q., and Chen, Z. Cross-Domain Sentiment Classification via Spectral Feature Alignment. In proc. of the 19th international conference on World wide web, pp. 751–760, 2010.
12. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., and Daelemans, W. Overview of the 2nd Author Profiling Task at PAN 2014. Proc. of the Conference and Labs of the Evaluation Forum (Working Notes), 2014.
13. Villegas, M. P., Garciarena Ucelay, M. J., Errecalde, M. L., and Cagnina, L. C. A Spanish text corpus for the author profiling task. In Proc. of XX CACIC. San Justo, Buenos Aires, Argentina, pp 1-10, 2014.
14. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., and Inches, G. Overview of the Author Profiling Task at PAN 2013. Notebook Papers of CLEF, pp. 23–26, 2013.
15. López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., and Villatoro-Tello, E. Inaoe's Participation at PAN'13: Author Profiling Task. Notebook PAN at CLEF 2013, 2013.
16. Finn, A., Kushmerick, N., and Smyth, B. Genre classification and domain transfer for information filtering. In Advances in information retrieval, Springer, pp. 353–362, 2002.
17. WEKA. Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/
18. Feldman, R., and Sanger, J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.
19. Cavnar, W. B., and Trenkle, J. M. N-gram-based text categorization. Ann Arbor MI, 48113(2):161–175, 1994.
20. Manning, C. D., Raghavan, P., and Schütze, H. Introduction to Information Retrieval. Cambridge University Press, New York, USA, 2008.
21. Lin, J. Automatic author profiling of online chat logs. Doctoral thesis, Monterey, California. Naval Postgraduate School, 2007.
22. Fan, R.-E., Chang, K.-W., Hsieh, C.-J.,Wang, X.-R., and Lin, C.J. LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research, 9:1871–1874, 2008.
23. Fúnez, D., Cagnina, L., and Errecalde, M. Determinación de Género y Edad en Blogs en Español Mediante Enfoques Basados en Perfil. In XVIII CACIC, 2013.