A Spanish Text Corpus for the Author Profiling Task

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

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
email: {villegasmariapaula74, mjgarciarenaucelay, merrecalde, lcagnina}@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 its potential applications in security, crime and marketing, among others. One of the main difficulties in this field is the lack of reliable text collections (corpora) to train and test automatically derived classifiers, in particular in specific languages such as Spanish. Although some recent data sets were generated for the PAN competitions, these documents have a lot of "noise" that prevent researchers from obtaining more general conclusions about this task when more formal documents are used. In this context, this work proposes and describes *SpanText,* a data collection of formal texts in Spanish language which is, as far as we know, the first collection with these characteristics for the author profiling task. Besides, an experimental study is carried out where the difference in performance obtained with formal and informal texts is clearly established and opens interesting research lines to get a deeper understanding of the particularities that each type of documents poses to the author profiling task.

Keywords: Author Profiling, Natural Processing Language, Spanish Text Corpus

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 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]. This task of growing importance is a very active research area because of its utility in security, crime, marketing and business. For example, author profiling can help police officers to detect cyber-pedophiles [9, 10], or to determine the veracity of a suicidal letter, among others applications. From the point of view of business, a company may be interested in

knowing what kind of audience is interested in its products and which products customers prefer, analyzing the comments on their websites using opinion mining [6]. This task has been mainly addressed by using supervised machine learning approaches and, in consequence, a key aspect is the availability of (categorized) text corpora for training and testing the classifiers. Nevertheless, automatic approaches to generate those data collections become difficult because one must have access not only to the text but also to personal information (like age and gender) of the author, which in many cases might be false. To the best of our knowledge there are diverse manually or semi-manually built corpus in English [2, 5, 11, 12], Arabic [11], Dutch [13, 14], Vietnamese [15] and in Greek [16]. However, in Spanish, the only resource available for the author profiling task at this moment is the one provided at the PAN-2013 competition [17]. This is a large corpus (hundreds of thousands documents) which has a high level of "noise" due to the automatic recovering of documents from the Web and the intrinsic characteristics of content generated in the Web 2.0: semantically incorrect (or incomprehensible) phrases, spelling, syntactical and grammatical mistakes and evolving, non-standard vocabulary (slangs, dialects and neologisms to shorten words), among others.

The above mentioned characteristics make this collection attractive to study the author profiling problem with informal, user-generated content on the Web. However, it does not correctly reflect a lot of "formal" information that could be used in an author profiling task such as those available in news, reports, scientific articles, etc. This lack of "formal" documents for the (Spanish) author profiling task not only leaves out of consideration much information that could be useful in specific tasks; it also prevents researchers from analyzing what are the differences and similarities of the author profiling task on domains with different level of "noise" (formality). As it has been observed in previous works on (topic) text classification [18] this can be an aspect that directly affects the performance of the classifier.

In this paper we introduce a Spanish text corpus for the Author Profiling problem named *SpanText*. This proposed corpus was manually collected and labeled resulting in a collection of 1000 formal documents about different topics written by women and men of different ages. Besides categorize the genre of the authors of each document, we classify the ages in three groups according to the categories considered in PAN-2013 competition task: 10's, 20's and 30's. We present a balanced version of this corpus in which the amount of documents in all the categories is similar. The unbalanced version has different amount of document in each category which corresponds to the same proportion of documents in the corpus of PAN-2013. Moreover, some initial experiments with well-known text representations and algorithms are presented which clearly show the differences in performance obtained with informal and formal documents.

The rest of this article is organized as follows: in Section 2, we briefly introduce the Author Profiling task. In Section 3, the main characteristics of the corpus and the data collection process are analyzed. Section 4 describes an experimental study comparing the results of standard classifiers with the proposed corpus versus the results obtained with a sub-collection of the PAN-2013's corpus. Finally, in Section 5 some conclusions are drawn and future works are proposed.

2 Author Profiling Task

Nowadays, the increasing use of online social networks like Facebook and Twitter has made available a huge amount of information in plain text. Such information can be used to infer important information of the author profile of a text [19]. The Author Profiling task (APT) consists in knowing as much as possible about an unknown author, just by analyzing the given text [20]. Thus, for example, profiling is used to determine an author's gender, age, level of education, geographic origin, native language and personality type [21].

The APT has mainly focused on documents written in English and, according to our knowledge, there is an important lack of available resources for the Spanish language. This situation has started to change in the last year with the papers presented at the PAN-PC-2013 [17, 19]. The organizers of PAN-2013 considered the gender and age aspects of the author profiling problem, both in English and Spanish languages, because these are two of the most spoken languages in the world [21].

The PAN-2013 corpus includes texts of blogs and other social media with special emphasis on the use of everyday language [21]. This kind of texts allows researchers to study the personality of people and the social processes. Thus, the texts are informal and many of them include typos, images, hyperlinks, emoticons, contractions, etc. These contents inside the texts can be considered as a kind of "noise" for a classifier if they are not correctly dealt with. In the case of the PAN-2013 competition, the training and testing corpora consisted of 75900 and 8160 documents respectively and both were balanced in terms of gender but not considering the age.

The gender classification was a binary task (female versus male) and by age, three classes were considered: 10s (people around 13 to 17 years old), 20s (people around 23 to 27 years old) and 30s (people around 33 to 47 years old) [3, 21]. The characteristics of texts and the process used to gather them became this "noisy" corpus a very challenging data set for any classifier. However, from the results of the competition it can be seen that some approaches like the one used in [20] (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 with other documents in Spanish that do not exhibit such a high level of "noise" and informal writings. For English language some studies with formal documents were presented in [3, 8] but similar studies are not possible in Spanish due to the lack of a corpus without noise and including formal documents. That was the main reason that motivated us to propose *SpanText*, a corpus with the desired characteristics.

3 Data

SpanText is a set of "formal" Spanish documents extracted from the Web that were written by different authors. In this context, we will use the term "formal" 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. In other words, we are focusing on the kind of texts

that one is supposed to find in newspapers, reports of students, books, etc. After the collection process and, due to the divergence in character set encodings of the gathered texts (ASCII, UTF-8, ANSI and ISO-8859-1), the format of documents was unified by converting them to the same codification (UTF-8 encoding). Two versions of the *SpanText* corpus are presented. The balanced version has a similar number of documents in each category. The unbalanced version has different amount of documents in the categories but the number is proportional to the one corresponding to the Spanish corpus of PAN-2013. A detailed analysis of the characteristics in each version is presented below.

3.1 Data Collection

The balanced and unbalanced versions of *SpanText* corpus have the following characteristics:

- Each corpus includes 1000 documents in Spanish language collected from the Web.
- Texts in both corpora were written by Spanish speakers from Spain and different Latin American countries.
- Texts "speak" about different topics.
- Each single text was guaranteed to have at least 150 words.
- All documents include "formal" text, i.e. without (or with a minimum amount of) typos, contractions, slangs, hyperlinks, labels, graphics, figures and emoticons.
- There is one document (file) per author but each file may contain two or more articles written by this author.

We addressed the same basic demographic information on authors considered at PAN-2013: age and gender. All the documents are labeled both for age and gender and were obtained from trustworthy online blogs and newspapers in which this kind of information was available. For age detection, we considered three classes: 10s (8-19), 20s (20-29) and 30s (30+). Although these age ranges are slightly different from those proposed in PAN-2013, we believed that ours are more realistic.

For the 20s and 30s classes we consider mostly texts from newspapers and blogs. Additionally, we provide a wide spectrum of topics, making the task of determining age and gender more colloquial than in PAN-2013 collection. For the 10s class, the collected documents consist mainly of stories written by children for literary competences and school projects, because these were easier to find.

After performing the collecting process, all the documents were manually labeled considering the information provided by the Web site or the data revealed inside the text.

In the balanced version of the *SpanText* the documents are uniformly distributed in the categories while in the unbalanced version, the number of documents is proportional to the amount of documents of PAN-2013 corpus. Table 1 shows the distribution of documents in both versions of *SpanText*.

Table 1. Documents distribution of the balanced versus unbalanced version of *SpanText* corpus considering genre and age categories.

Version	Female	Male	10's	20's	30's
Balanced	500	500	333	333	334
Unbalanced	300	700	200	300	500

3.2 Analysis

Table 2 shows the main characteristics of both versions of the *SpanText* corpus. It includes the total number of words, sentences, number of unique words found in all documents (Vocabulary) and the size of each version (in Kilobytes).

Table 2. Characteristics of *SpanText* corpus.

Version	Authors (Docs)	Words	Sentences	Vocabulary	Size
Balanced	1000	792567	36 116	53384	6 540
Unbalanced	1000	814031	36 781	54 648	6708

The characteristics of both versions are very similar and, in the unbalanced version (see Table 1), we can observe that there are more male authors than female authors.

Figure 1 shows the distribution of words and sentences in the balanced *SpanText* corpus. On the left side of Fig. 1, we can observe that most of documents have less than 2000 words. In fact, the average number of words per documents is 792 with an important standard deviation of 512. Only a few documents have over 2500 words. With respect to the number of sentences (right side of Fig. 1) we can observe that most of documents have less than 50 sentences (36 in average). Around 200 documents have over 50 sentences and only one document has more than 250 sentences (451 sentences).

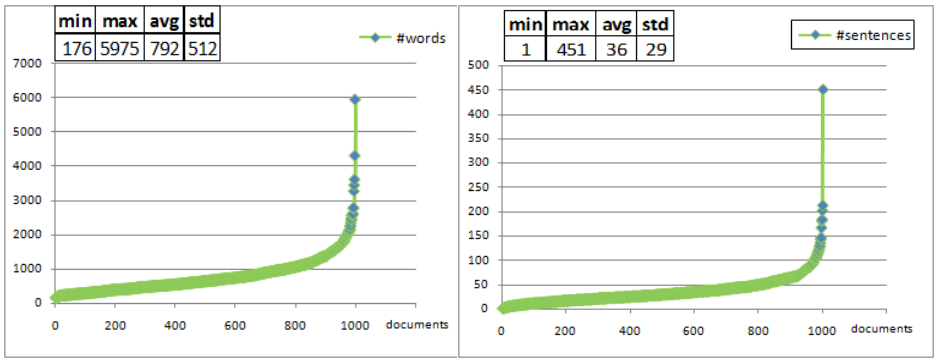

Figure 1. Distribution of words and sentences of the balanced *SpanText* corpus.

Figure 2 shows the distribution of words and sentences in the unbalanced *SpanText* corpus. On the left side, we can observe a similar situation that in the balanced version, that is, most of documents have less than 2000 words with an average number of words per documents of 814. Only two documents have over 4000 words. With respect to the number of sentences (right side of Fig. 2) we can observe that most of documents have less than 50 sentences and only very few documents have over 100 sentences (only two documents have more than 150 sentences).

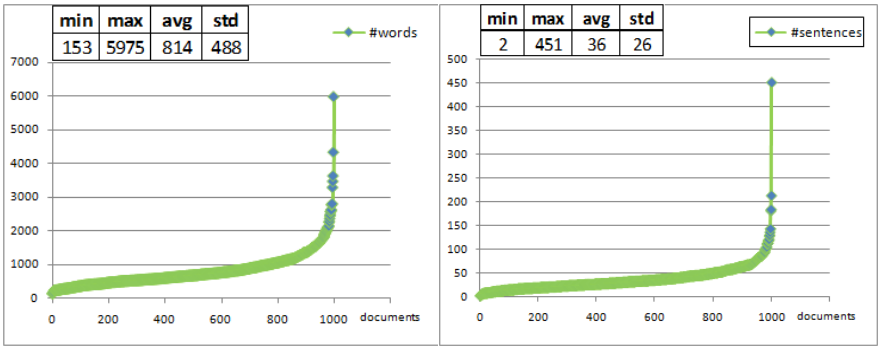

Figure 2. Distribution of words and sentences of the unbalanced *SpanText* corpus.

3.3 Classes

From the combination of the gender and age of the author, six classes can be derived. These were identified as: 10smal, 10sfem, 20smal, 20sfem, 30smal, 30sfem. In Tables 3 and 4, statistics of the balanced and unbalanced versions of *SpanText* for each class are shown. They include, for each age-gender combination (category), the number of authors (or number of documents), total number of words, average number of words and sentences respectively, vocabulary (number of different words) and size (in KiloBytes).

From Table 3 we can observe that in the balanced version, male authors use more words, more sentences and have a richer vocabulary than female authors, except for the 30's class. Women around thirty's tend to use significantly more words and more sentences (almost double) than men.

In Table 4 (the unbalanced version), conclusions are similar to those obtained from the balanced version. Averaged values show similar proportions between men and women to the ones shown in Table 3. However, the number of words and vocabulary seem to be a little more balanced due probably to a larger number of male authors (350 vs 150 documents for the "30s" category).

Table 3. Distribution of documents per class for the balanced version of *SpanText*.

Category	Authors (Docs)	Words	Average words per doc	Average sentences per doc	Vocabulary	Size
10smal	166	103722	624	29	13542	590
10sfem	167	82057	491	28	11999	474
20smal	167	120462	721	33	17302	700
20sfem	166	116874	704	27	16304	686
30smal	167	132829	795	36	19758	797
30sfem	167	236623	1417	60	27103	1429

Table 4. Distribution of documents per class for the unbalanced version of *SpanText*.

Category	Authors (Docs)	Words	Average words per doc	Average sentences per doc	Vocabulary	Size
10smal	140	87392	624	29	12066	497
10sfem	60	30444	507	27	6291	176
20smal	210	151143	719	34	20176	882
20sfem	90	70176	779	31	12015	423
30smal	350	259502	741	33	29151	1566
30sfem	150	215374	1435	60	25651	1303

4 Experimental Study

In this section, we compare the performance of basic approaches to solve the author profiling problem when a formal corpus as *SpanText* and an informal corpus as PAN-2013 are used. This study aims at showing how the level of formality (or conversely the presence of noise) affects the performance of the classifiers. Although we do not try to establish other deeper relationships about the author profiling task on both types of documents, this study will offer evidence of the need for paying particular attention to this key aspect in the APT with documents in Spanish language.

Due to the huge size of the Spanish corpus used at PAN-2013, we have generated (smaller) sub-corpora from that collection in order to make comparable the experimental results. The balanced and unbalanced versions obtained in this case have the same distributions of documents as the corresponding versions of *SpanText.* In this case, documents were randomly selected from PAN-2013 Spanish corpus to reflect the different types of (informal) documents that we could expect to find in a real situation. Table 5 shows the characteristics of these versions of the sub-corpus. Note that the number of words, sentences and vocabulary is quite different compared with those of *SpanText,* mainly due to the short length of posts of blogs collected when the PAN-2013 was automatically generated.

Besides controlling that formal and informal corpora have the same document distributions, a preprocessing stage was applied to remove some noise (like quotes, links, smiles, HTML tags, etc.) present in the sub-corpus of PAN-2013.

Table 5. Distribution of documents per class for the balanced and unbalanced versions of sub-corpus obtained from the Spanish PAN-2013 collection.

Category		Authors (Docs)	Words	Average words per doc	Average sentences per doc	Vocabulary
Balanced	10smal	166	25428	153	9	6661
	10sfem	167	21758	130	11	5039
	20smal	167	40034	239	14	8486
	20sfem	166	39221	236	29	8065
	30smal	167	44636	267	19	8731
	30sfem	167	40182	240	21	8097
Unbalanced	10smal	140	22303	159	9	5942
	10sfem	60	5440	90	6	1691
	20smal	210	52774	251	15	10425
	20sfem	90	25984	288	42	5772
	30smal	350	94016	268	17	15231
	30sfem	150	36914	246	22	7536

Regarding the representation of documents, we used the standard *bag of words* (BOW) [3, 13, 14, 19, 20] and character 3-grams representations [1, 14, 20, 22] with TF-IDF and Boolean (binary) weighting [23]. We also used well-known machine learning algorithms to train the classifiers: LibLINEAR [4, 13, 14, 20, 22] and Naïve Bayes [1, 22, 23]. The accuracy of the algorithms was determined by using a 10-fold cross-validation approach. As implementation of these algorithms, we used those available in the WEKA data mining software [24] and a Python script was programmed to obtain the character 3-grams of the documents. The percentages of correctly classified instances (accuracy) obtained with each version of *SpanText* and the sub-corpus of Spanish PAN-2013 (SCPAN-13) are shown in Table 6.

Table 6. Percentage of correctly classified instances using 10-fold cross-validation with LIBLINEAR/NaïveBayes algorithms.

Instance	TF-IDF words	TF-IDF 3grams	Boolean words	Boolean 3grams
SpanText balanced	**52.9/49.6**	**43.4/48.7**	**52.7/54.7**	**45.5/47.1**
SCPAN-13 balanced	24.7/24.3	26.3/24.8	24.9/20.6	24.7/20.4
SpanText unbalanced	**58.4/55.5**	**47.2/50.9**	**55.2/61.2**	**49.0/51.2**
SCPAN-13 unbalanced	30.7/22.5	30.6/27.8	32.1/17.5	30.6/12.3

The highest accuracy values obtained with each representation and version of the corpora are highlighted in boldface in Table 6. It is worth noting that the percentages of correctly classified instances of *SpanText* are clearly higher than the percentages obtained with the sub-corpus of the PAN-2013 competition. Particularly when the TF-IDF words representation was used, the percentages are almost duplicated. A similar pattern can be seen in the remaining experimental instances. We think that these results are indicative of the significant impact that different levels of noise have on the author profiling task in Spanish documents. That motivates us to use the "formal"

data set proposed in the present article in more elaborated studies that will be described in the future work.

5 Conclusions and Future Work

The problem of identifying the author profile from a text has received growing attention mainly due to its potential applications in areas like forensics and marketing, among others. However, a considerable problem for the development of solutions, in particular when working with documents in Spanish language, is the lack of reliable data sets to train and test automatically generated classifiers.

In this context, this paper makes an interesting contribution by presenting *SpanText*, a data collection of formal texts in Spanish language which is, as far as we know, the first collection with this characteristics proposed for the author profiling task with documents in Spanish. This corpus is available for other researchers interested in this field by e-mailing to any of the article's authors. Besides, as a secondary contribution, an initial experimental study was carried out where the difference in performance obtained with formal and informal texts are clearly established. Also in this case, and to the best of our knowledge, there are no previous similar comparative studies with Spanish documents.

Even though this experimental study is only an initial analysis of the difficulties that each kind of corpus poses to traditional machine learning approaches and document representations, it opens interesting research lines to be addressed in the future. First of all, it would be interesting analyzing if the same document representation that was used in the winning approach in the PAN-2013 competition obtains such a good performance in formal documents in comparison to other simpler approaches likes the ones considered in the present article. In this context, a detailed analysis could be carried out about how robust are the different features when applied to documents with different level of noise.

SpanText also will allow studying what are the stylometric and lexical features (common to both kinds of documents) that are most informative to discriminate among the different categories. That will serve to identify useful features and patterns that are kept unalterable independently of the level of informality of the documents. Besides, analyzing which are the more informative features in each case (for instance, by using simple techniques like information gain) much information could be obtained if some kind of "normalization" (transformation for noise elimination) is applied in the future on the informal texts in order to obtain better results.

References

1. Koppel, M., Argamon, S., and Shimoni, A. R. Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing. Volume 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, volume 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 Proceedings 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. Proceedings 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. Volume 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. Volume 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. Escalante, H. J., Villatoro-Tello, E., Juárez-González, A., Villaseñor-Pineda, L., and Montes-y-Gómez, M. Sexual Predator Detection in Chats with Chained Classifiers. 4th Workshop on Comp. Approaches to Subjectivity, Sentiment, & Social Media Analysis, NAACL-HLT 2013, pp. 46–54, 2013.
10. Villatoro-Tello, E., Juárez-Gonzalez, A., Escalante, H. J., Montes-y-Gómez, M., and Villaseñor-Pineda, L. A two-step Approach for Effective Detection of Misbehaving Users in Chats. In Working Notes of the CLEF 2012, 2012.
11. Estival, D., Tanja,G., Hutchinson,B., Pham, S. B., and Radford,W. Author Profiling for English and Arabic Emails. Natural Language Engineering, Cambridge U. Press, 1998.
12. Estival, D., Gaustad, T., Pham, S. B., Radford, W., and Hutchinson, B. Author Profiling for English Emails. In 10th Conf. of the Pacific Association for Computational Linguistics, pp. 263–272, 2007.
13. Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. "How Old Do You Think I Am?": A Study of Language and Age in Twitter. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, 2013.
14. Peersman, C., Daelemans, W., and Van Vaerenbergh, L. Predicting Age and Gender in Online Social Networks. In Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. ACM, pp. 37–44, 2011.
15. Pham, D. D., Tran, G. B., and Pham, S. B. Author Profiling for Vietnamese Blogs. In Asian Language Processing, 2009. IALP '09. International Conf. on. IEEE, pp. 190–194, 2009.
16. Mikros, G. K. Authorship Attribution and Gender Identification in Greeks Blogs. Methods and Applications of Quantitative Linguistics, pp. 21, April 2012.
17. 9th Evaluation Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2013). http://pan.webis.de/, 2013.
18. Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., Tserpes, K. Representation Models for Text Classification: A Comparative Analysis Over Three Web Document Types. WISM 2012.
19. 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.
20. 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.
21. 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.
22. Lex, E., Juffinger, A., and Granitzer, M. A Comparison of Stylometric and Lexical Features for Web Genre Classification and Emotion Classification in Blogs. In 2010 Workshop on Database and Expert Systems Applications (DEXA), IEEE, pp. 10–14, 2010.
23. Witten, I. H., and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 3rd edition, 2011.
24. WEKA. Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/