Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 959–964

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS)
April 6-9, 2020, Warsaw, Poland

A method for detecting the profile of an author

Jesus Silva[a]*, Silvia García[b], María Alejandra Binda[c], Fredy Marin Gonzalez[d], Rosio Barrios[e], Bellanit Leon Castro[f], Ligia Castro[g]

[a, b,c] *Facultad de Negocios, Universidad Peruana de Ciencias Aplicadas, Lima, Peru*
[d, g] *Universidad de la Costa, Barranquilla, Colombia*
[e] *Corporacion Universitaria Minuto de Dios (UNIMINUTO), Barranquilla, Colombia.*
[f] *Corporacion Universitaria Latinoamericana (CUL), Barranquilla, Colombia.*

Abstract

This paper presents a method for detecting an author's profile using the following two elements: gender and age. This is based on a set of dialogues, written in two languages: English and Spanish, provided for Author Profiling competence within the evaluation forum "Uncovering Plagiarism, Authorship, and Social Software Misuse" (PAN2018). Counts of lexical, semantic, and syntactic characteristics are used to generate a two-phase classification system, which first classifies gender and then age. The results obtained show that, with the amount of data available, it is possible to characterize both the age and gender of an author with an accuracy greater than 50%. However, these values could be improved by having more evidence of information in the training data.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the Conference Program Chairs.

Keywords: Supervised Classification, PAN 2018, Gender, Age, Random forest.

1. Introduction

The detection of an author's profile is an increasingly important problem in various fields of knowledge such as

* Corresponding author. Tel.: +51-975857503.
E-mail address: jesussilvaUPC@gmail.com

1877-0509 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2020.03.101

forensic medicine, security and marketing. For example, from the perspective of forensic linguistics, it would be important to know the linguistic profile of the author of a harassing message. Similarly, from a marketing point of view, companies may be interested in knowing, through the analysis of online blogs and product reviews, what kind of people comment on their products, and thus direct their advertising campaigns towards a certain gender or age range [1]. A detailed description of the subject's background can be found at:[2][3][4][5][6][7][8][9][10][11].

This paper proposes a method for detecting two aspects of the profile of authors in chats or blogs: gender and age. This method has made it possible to create a system that, based on a set of dialogues written by different people (each dialogue contains the gender and age group of the person who wrote it), to catalogue a new set of dialogues (evaluation set) and to determine the two aspects of the profile mentioned above. The texts used are presented in English and Spanish.

2. The Method

The training set consists of XML-type documents containing conversations on different topics grouped by author and tagged with their language (English or Spanish), gender (male and female) and age group. There are three age groups:

1. 10s: People from 13 to 17 years old.
2. 20s: People from 23 to 27 years old.
3. 30s: People from 33 to 47 years old.

The English corpus contains 236,600 authors, while the Spanish corpus contains 75,900. Table 1 shows some statistics of the dialogues, separated by language and gender [12].

Table 1. training set for the Author profiling task.

	English		**Spanish**	
	Female	**Male**	**Female**	**Male**
Total of dialogues	118,300	118,300	37,950	37,950
Dialogues de “10s”	8,600	8,600	1,250	1,250
Dialogues de “20s”	42,900	42,900	21,300	21,300
Dialogues de “30s”	66,800	66,800	15,400	15,400
Vocabulary	1,228,711	1,219,020	533,873	592,605
Average of words	798	705	249	273
Dialogues with more words	10,648	12,917	11,806	11,714

The vocabulary is very extensive, especially for the English language, which is logical since there are so many dialogues for this language. In the case of the Spanish language, the vocabulary is more extensive in the male category, while for the English language, the vocabulary is more extensive in the female category. In addition, it is observed that in all age categories, the number of texts is the same for both genders (in this sense, it is a gender-balanced corpus for each age range) [13].

Analyzing the sets of texts, it was concluded that they present many misspelled words, truncated words, emoticons and vocabulary from blogs and chats. Therefore, the texts are expanded by building some lexical resources for both languages, such as: a dictionary of emoticons, a dictionary of abbreviations, a dictionary of common words in SMS and the most used contractions. Using the previously mentioned lexical resources, each occurrence of emoticons and contractions could be substituted in the training corpus by its corresponding meaning. Punctuation marks and non-printable characters were also eliminated. After applying this pre-processing, vocabulary in both languages was drastically reduced.

In the phase of extraction of characteristics, most of the researches carried out use all the vocabulary, however, in this case it is observed that this type of approach would consume too much space and time resources of computation, and even certain automatic learning tools would not be able to support such a volume of information, that is why it is proposed to use the following counts:

1. Grammatical Categories: The grammatical category of each work was obtained within the texts, to later carry out the count of each of them. For this, the tool Tree-Tagger [14] was used for the Spanish language and the Stanford POS-tagger [15] for the English language, obtaining 102 characteristics for the Spanish language and 52 for the English language.
2. Closed Words: Within this category, groups of words are classified as prepositions, conjunctions and determinants. Thus, each closed word represents a characteristic, and its value in each instance is given by the number of times it appears in the conversation. 195 words are obtained for the English language and 178 for the Spanish language.
3. Suffixes: The existing suffixes for both languages were taken as characteristics. As in the previous sets, each suffix represents a characteristic, and the number of times it appears in each conversation is the value for that attribute. In this set, 131 characteristics were obtained for the English language and 172 for the Spanish language.
4. Signs: All existing punctuation marks are counted.

As mentioned above, this article uses the automatic learning approach to detect the profile of authors. This approach starts from the premise of the existence of a supervised corpus that is used to train a classification model, which is then used to calculate the class associated with an input text whose class is unknown. Using the above-mentioned characteristics, the following classification models were generated using the Random Forest method [16]:

1. Gender: All texts from the training set are used to classify gender (male, female), as a classifying attribute.
2. AgeMale: All texts written by men are grouped using the age range (10s, 20s and 30s), as a classifying attribute.
3. AgeFemale: All texts written by women are grouped using the age range (10s, 20s and 30s), as a classifying attribute.

3. Results

This section presents the results obtained for both cross validation and software submission evaluation. For the analysis of the obtained results, precision metrics, recall and the F1 [18][19][20][21] measure are used, which is defined as the harmonic mean between precision and recall.

3.1 10-fold cross validation

As mentioned above, only the training dataset is used in this type of assessment. Table 2 shows the results obtained by classifying only gender, while Table 3 shows the values obtained by classifying only the author's age. The F1 measurement value obtained in the case of the gender classification is 0.5541 for the Spanish language, while it is 0.5654 for the English language. The age classification showed an F1 equal to 0.3854 for the Spanish language and 0.4825 for the English language. As it can be observed, on the data set used as training, the average values validate the hypothesis raised, which indicates that it is easier to detect the author's gender than his age. The behavior observed was similar for both languages (Spanish and English).

Table 2. Evaluation by cross validation using gender (male and female) as classifying attribute.

Category	Spanish			English		
	Precision	Recall	F1	Precision	Recall	F1
Male	0.5412	0.6547	0.5987	0.5398	0.6654	0.5995
Female	0.5654	0.4514	0.5095	0.5754	0.4654	0.5313
Average	0.5533	0.5531	0.5541	0.5576	0.5654	0.5654

However, this behavior could be affected in some way by the number of samples for the class of authors that are in the range of 10 years (10s), since the few may not correctly represent the class and generate an over-adjusted model that does not behave adequately on the corpus of evidence. This class is too small in relation to the other two

(20s and 30s) and therefore, the classification process may tend to favor the classes with the largest number of samples. Under this reasoning, and considering a classification process based on two phases, two evaluations are proposed: first to classify the author's gender followed by the age classification (as proposed in the initial methodology) and in a second evaluation the inverse process is carried out.

Table 4 presents the precision, recall and F1 values for each classification model applied in the second phase of the initial method, i.e., taking the gender models to determine age.

Table 3. Cross validation assessment using age (10s, 20s and 30s) as the classifying attribute

Category	**Spanish**			**English**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
10s	0.0000	0.0000	0.0000	0.3345	0.0321	0.0478
20s	0.7401	0.7854	0.7021	0.5245	0.5012	0.4985
30s	0.5541	0.3521	0.4541	0.6584	0.7578	0.9012
Average	0.4314	0.3792	0.3854	0.5058	0.4304	0.4825

Table 4. Results of models that use age (10s, 20s and 30s) as a classifying attribute in the second phase of classification.

Model	**Category**	**Spanish**			**English**		
		Precision	Recall	F1	Precision	Recall	**F1**
Age	10s	0.1425	0.0042	0.0079	0.3201	0.0298	0.0521
Male	20s	0.6214	0.7957	0.6985	0.5214	0.4845	0.4995
	30s	0.5298	0.3654	0.4341	0.6365	0.7471	0.6854
	Average	0.4321	0.3954	0.4120	0.4987	0.4214	0.4541
Age	10s	0.0000	0.0000	0.0000	0.3124	0.0301	0.0552
Female	20s	0.6154	0.7754	0.6852	0.5210	0.4654	0.4965
	30s	0.5241	0.3654	0.4214	0.6586	0.7752	0.7021
	Average	0.3754	0.3802	0.3852	0.4954	0.4214	0.4548
General Average		0.4001	0.3798	0.3899	0.4952	0.4214	0.4512

A better behavior is observed in the age range of 20 years, for dialogues written in Spanish (regardless of gender). In the case of dialogues written in English, the best result is obtained in the age range of 30 years. From this particular point of view, the results are co-related to the number of dialogues in the training dataset by age range. The experiment yields an F1 of 0.39 for Spanish and 0.45 for English. As expected, these F1 values are lower than those obtained in the first qualifying phase. Badly classified dialogues in phase one are destined to cause a margin of error close to 45 % in the second qualifying phase.

In the case of the second evaluation, a classification model was created in which only age is used as the classifying attribute. The output of this model can be "10s", "20s" or "30s" (results in table 3). Once the author's age range was identified, the gender is identified, according to the age identified in the previous phase. Thus, there are now three additional classification models, one that trains on the dialogues written by people in the age range of 10 years, another on the age range of 20 years, and the last on the age range of 30 years.

The final output indicates the age range and the identified gender to which a given author belongs. Although this system has the same input and output as the one presented above, the results shown in Table 5 show the impact of selecting one phase over the other. The values of precision, recall and F1 for each classification model applied in the second phase are presented again.

When comparing the results obtained with respect to the previous scheme, a greater loss of precision is observed in all age ranges, regardless of language. In particular, in the case of the Spanish language, it was not possible to distinguish the dialogues written by the authors in the age range of 10 years. This fact again suggests that the results are correlated with the quality of the training corpus, in such a way that the dialogues are not representative of the class and may have generated an over-adjusted model, as mentioned above.

Finally, it was decided to investigate the behavior of the classification process using a single model, i.e., a single phase in which there are 6 different classes associated with the different categories presented by the corpus: male in the range of 10 years, male in the range of 20 years, male in the range of 30 years, female in the range of 10 years, and so on. The results obtained and their comparison with the two models previously presented (Gender->Age and Age->Gender) are shown in Figure 1.

Table 5. Results of models that use gender (male and female) as a classifying attribute in the second phase of classification.

Model	**Category**	**Spanish**			**English**		
		Precision	Recall	F1	Precision	Recall	**F1**
Gender 10s	Male	0.0000	0.0000	0.0000	0.2014	0.2524	0.2214
	Female	0.0000	0.0000	0.0000	0.3035	0.2452	0.2665
Gender 20s	Male	0.3332	0.4013	0.3645	0.3456	0.3965	0.3785
	Female	0.3542	0.2785	0.3012	0.2963	0.2325	0.2621
Gender 30s	Male	0.2845	0.3458	0.3024	0.3214	0.3854	0.3521
	Female	0.2985	0.2214	0.2456	0.3785	0.3123	0.3456
General Average		0.2123	0.2123	0.2132	0.3123	0.3021	0.3021

Using the 6 classes avoids dragging errors from one phase to another, however, there is the problem of the increase in the number of classes. The classifier has more difficulty to discern between the different possibilities. Basically, the best approximation that could be had in the experiments presented in this paper was when a two-phase classification process was used, identifying first the gender and then the author's age range. Based on these results, it was decided to use this approach for the evaluation of the test corpus in the software-submission approach.

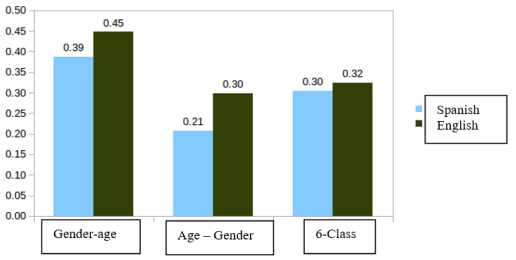

Fig. 1. Comparison between the approximations of two phases presented and a single classifier with 6 classes.

3.2 Software submission

The system presented for the competition "PAN 2018" (Aleman) ranked 7 with an accuracy of 55.24 % for gender detection, 59.32 % for age detection and an overall accuracy of 32.14 %. These results are quite similar to those obtained in the tests with the training set and although overall the accuracy does not exceed 50%, no participating team reached this value, being first place with 39.56% accuracy.

This approach was evaluated within the framework of competition in order to detect conversations involving pedophiles. The results obtained place us in 12th place out of 21 participating teams. In fact, there are 8 teams that detect 100% the gender of the predators, however, this result is not as significant since it is well known that most sexual predators are male, which is also reflected in the data set of the competition. This approach obtained 74% accuracy in gender detection, which is consistent with the results reported throughout this paper. The conclusion is that the set of characteristics should be refined according to the type of classifying attribute, since it is not the same to classify gender as age. Gender is influenced by the character of people and the communication habits between men and women, but in the case of age, the size of the vocabulary, for example, should be significant, since people tend to increase and modify the vocabulary as they grow up.

4. Conclusions

This article presented a method for gender and age detection in blogs and chats. This methodology uses counts of lexical, syntactic, and semantic characteristics to represent people's dialogues in order to train a supervised classification model to determine an author's gender and age range.

The results obtained using only the training set show that there is a better classification in gender than in age, however, in neither case is exceeded 55 % of F1 measure, in addition, this measure lowers when the two classifiers are joined, reaching a F1 value between 40 % and 44 %.

In particular, the number of dialogues for authors in the 10-year age range is very low and therefore, there is a risk that they are not representative of class in real life. Thus, the quality of the corpus had a negative effect on the classification process, over-adjusting the training data and generating a model that cannot adequately recognize the test data, especially for the set of dialogues that were written by authors in the 10-year age range.

References

[1] Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006. (2006) 199–205

[2] Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2) (February 2009) 119–123

[3] Peersman, C., Daelemans, W., 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. SMUC '11, New York, NY, USA, ACM (2011) 37–44

[4]. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "how old do you think i am?": A study of language and age in twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. ICWSM 2013 (2013)

[5] Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. In: Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science (NLPCS-2013). (2013)

[6] Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Procee- dings of the International Conference on New Methods in Language Processing, Manchester, UK (1994)

[7] Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Human Language Technology Conference (HLT-NAACL 2003). (2003)

[8]. Viloria A., Lis-Gutiérrez JP., Gaitán-Angulo M., Godoy A.R.M., Moreno G.C., Kamatkar S.J. (2018) Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching - Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[9] De Werra, D.: An introduction to timetabling. European Journal of Operational Research, vol. 19, no. 2, pp. 151–162 (1985).

[10]Obit, J. H., Ouelhadj, D., Landa-Silva, D., Vun, T. K., & Alfred, R.: Designing a multi- agent approach system for distributed course timetabling, pp. 103–108, doi:10.1109/HIS.2011.6122088 (2011)

[11]Lewis, M. R. R.: Metaheuristics for university course timetabling. Ph.D. Thesis, Napier University (2006)

[12].Deng, X., Zhang, Y., Kang, B., Wu, J., Sun, X., & Deng, Y.: An application of genetic al- gorithm for university course timetabling problem, pp. 2119–2122, doi:10.1109/CCDC.2011.5968555 (2011)

[13].Mahiba, A.A. & Durai, C.A.D.: Genetic algorithm with search bank strategies for universi- ty course timetabling problem. Procedia Engineering, vol. 38, pp. 253–263 (2012)

[14].Soria-Alcaraz, J. A.; Carpio, J. M.; Puga, Hé.; Melin, P.; Terashima-Marn, H.; Reyes, L.

[15]C. & Sotelo-Figueroa, M. A. Castillo, O.; Melin, P.; Pedrycz, W. & Kacprzyk, J.: Generic Memetic Algorithm for Course Timetabling. In: ITC2007 Recent Advances on Hybrid Approaches for Designing Intelligent Systems, Springer, vol. 547, pp. 481–492 (2014)

[16].Nguyen, K., Lu, T., Le, T., & Tran, N.: Memetic algorithm for a university course timeta- bling problem. pp. 67–71, doi:10.1007/978-3-642-25899-2_10 (2011)

[17].Aladag, C., & Hocaoglu, G.: A tabu search algorithm to solve a course timetabling prob- lem. Hacettepe journal of mathematics and statistics, pp. 53–64 (2007)

[18]Moscato, P.: On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms. Caltech Concurrent Computation Program (report 826) (1989).

[19] Viloria, Amelec; Lezama, Omar Bonerge Pineda. Improvements for Determining the Number of Clusters in k-Means for Innovation Databases in SMEs. Procedia Computer Science, 2019, vol. 151, p. 1201-1206.

[20] Kamatkar, S. J., Kamble, A., Viloria, A., Hernández-Fernández, L., & Cali, E. G. (2018, June). Database performance tuning and query optimization. In International Conference on Data Mining and Big Data (pp. 3-11). Springer, Cham.

[21] Viloria, Amelec, et al. Integration of Data Mining Techniques to PostgreSQL Database Manager System. Procedia Computer Science, 2019, vol. 155, p. 575-580.