On the Multilingual and Genre Robustness of EmoGraphs for Author Profiling in Social Media*

Francisco Rangel[1,2] and Paolo Rosso[1]

[1] NLE Lab, Universitat Politècnica de València, Spain
prosso@dsic.upv.es
[2] Autoritas Consulting, S.A., Spain
francisco.rangel@autoritas.es

Abstract. Author profiling aims at identifying different traits such as age and gender of an author on the basis of her writings. We propose the novel EmoGraph graph-based approach where morphosyntactic categories are enriched with semantic and affective information. In this work we focus on testing the robustness of EmoGraphs when applied to age and gender identification. Results with PAN-AP-14 corpus show the competitiveness of the representation over genres and languages. Finally, some interesting insights are shown, for example with topic and emotion bounded genres such as hotel reviews.

Keywords: Author profiling, Age Identification, Gender Identification, Emotion-labeled Graphs, EmoGraph

1 Introduction

Author profiling aims at identifying different traits such as age and gender of an author on the basis of her writings. Profiling an author is very important from a forensic and security viewpoint due to the possibility of profiling possible delinquents as well as from a marketing perspective due to the possibility of improving users segmentation. The growing interest in age and gender identification is notable in the scientific community. A shared task on author profiling has been organised at PAN Lab[3] of the CLEF initiative. The interest of PAN 2015 remains on identifying age and gender together with personality.

Pioneer investigations on author profiling were carried on by Pennebaker [4], who divided features into content and style-based. Similarly, in [1] authors approached the task of gender identification by combining function words with parts-of-speech (POS). A high variety of different approaches were used at PAN shared tasks [7, 6]. Participants used combinations of style-based features such as frequency of punctuation marks, capital letters, quotations, etc., joint POS tags and content-based features such as bag-of-words, TF-IDF of words, dictionary-based words, topic-based words, entropy-based words, or content-based features obtained with Latent Semantic Analysis. Few authors

* The research has been carried out in the framework of the European Commission WIQ-EI IRSES (no. 269180) and DIANA - Finding Hidden Knowledge in Texts (TIN2012-38603-C02) projects. The work of the first author was partially funded by Autoritas Consulting SA and by Spanish Ministry of Economics under grant ECOPORTUNITY IPT-2012-1220-430000.

[3] http://pan.webis.de

used emotions as features [3], but none of them focused on how users convey verbal emotions. Also, there are no investigations on graph-based representations to tackle author profiling. We approached the task of age and gender identification in Spanish with EmoGraphs in [5], obtaining competitive results and interesting insights on how people convey verbal emotions in their discourse. In this paper we are interested in investigating further the robustness of EmoGraphs from multilingual and genre perspectives in social media texts.

The rest of the paper is structured as follows. In Section 2 we introduce EmoGraphs. In Section 3 we explain the evaluation framework, presenting and discussing experimental results in Section 4. Finally, in Section 5 we draw some conclusions.

2 Emotion-labelled Graphs

Emotion-labeled Graphs (EmoGraphs) [5] obtains morphosyntactic categories with the Freeling library[4] for each word in all texts of an author. Each POS is modeled as a node in the graph and each edge defines a POS sequence in the text. The graph obtained is enriched with semantic and affective information. Adjectives, adverbs and verbs are annotated with their polarity and the Spanish Emotion Lexicon [8] and Wordnet Affect [9] are used to identify their associated emotions in Spanish and English respectively. WordNet Domains[5] is used to obtain the topics of nouns. On the basis of what was investigated in [2], verbs are annotated with one of the following semantic categories: i) perception (see, listen, smell...); ii) understanding (know, understand, think...); iii) doubt (doubt, ignore...); iv) language (tell, say, declare, speak...); v) emotion (feel, want, love...); vi) and will (must, forbid, allow...). We can see an example in Figure 1.

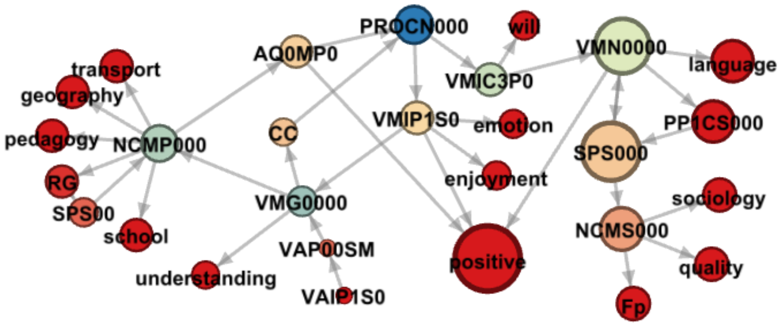

Fig. 1. EmoGraph of "He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público" (*"I have been taking online courses about valuable subjects that I enjoy studying and might help me to speak in public"*).

[4] http://nlp.lsi.upc.edu/freeling/

[5] http://wndomains.fbk.eu/

Once the graph is built, our objective is to use a machine learning approach to classify texts into the right age and gender. We obtain two kind of features on the basis of graph analysis: i) general properties of the graph describing the overall structure of the modelled texts, such as nodes-edges ratio, average degree, weighted average degree, diameter, density, modularity, cluster coefficient or average path length; ii) and specific properties of its nodes and how they are related to each other, such us eigenvector and betweenness values.

3 Evaluation Framework

In the following sections we describe the PAN-AP-14 corpus of the PAN Lab at CLEF and the methodology employed for identifying age and gender.

3.1 PAN-AP-14 Corpus

The PAN-AP-14 corpus incorporates four different genres: *i*) social media; *ii*) blogs; *iii*) Twitter; *iv*) and hotel reviews. The respective subcorpora cover English and Spanish, with the exception of the hotel reviews, which have been provided in English only. The author is labeled with age and gender information. For labeling age, the following classes were considered: *i*) 18-24; *ii*) 25-34; *iii*) 35-49; *iv*) 50-64; *v*) and 65+ . The number of different authors per genre and language in the test dataset is shown in Table 1. The dataset is balanced by gender. In the overview paper on the shared task [6] more details are given on how the different subcorpora were built.

Table 1. Distribution of the number of authors with respect to age classes per language in test set.

	Social Media		Blogs		Twitter		Reviews
	English	Spanish	English	Spanish	English	Spanish	English
18-24	680	150	10	4	12	4	74
25-34	900	180	24	12	56	26	200
35-49	980	138	32	26	58	46	200
50-64	790	70	10	10	26	12	200
65+	26	28	2	2	2	2	147
Σ	3376	566	78	54	154	90	821

3.2 Methodology

Our final representation is a combination of the presented EmoGraph with the 1,000 most frequent character 6-grams[6]. For training we used the training dataset of the PAN-AP-14 corpus. Several machine learning algorithms were evaluated with the training

[6] We combine our representation with n-grams due to the good results of other participants by using them in the task [6]. We selected n=6 due to experimental results on the training set.

set and selected the best ones: *i*) Simple logistic in Englih Twitter for gender identification; *ii*) Support Vector Machines in Spanish blogs, English reviews and English social media for both gender and age, and Spanish Twitter for age identification; *iii*) and AdaBoost with Decision Stump for all the rest. To compare our results with the ones of the participants in the PAN 2014 task, we evaluated our models on the test set.

4 Experimental Results

In this section results are presented and discussed. The first subsection presents the overall accuracy obtained in the task and compare them to the best method presented at PAN 2014 for each corpus and task. The second subsection shows the analysis of the impact of EmoGraphs.

4.1 Age and Gender Identification

As can be seen in Figure 2, results for Spanish are better than for English, except maybe in blogs. This may be due to the highest variety in the morphological information obtained with Freeling for both languages. The Eagles group[7] proposed a series of recommendations for the morphosyntactic annotation of corpora. Freeling obtains 247 different annotations for Spanish whereas it obtains 53 for English. For example, in the Spanish version[8] the word ”cursos” (courses) for the given example in Figure 1 is returned as NCMP000 where NC means common noun, M means male, P means plural, and 000 is a filling until 7 chars; in the English version, the word ”courses” is annotated as NNS.

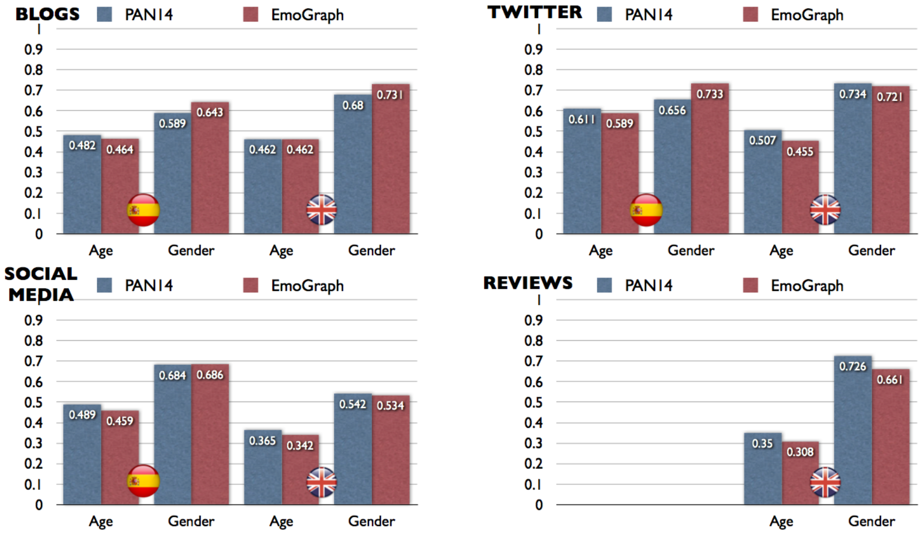

Fig. 2. Accuracies of the best PAN14 team vs. EmoGraph on different languages and genres.

[7] http://www.ilc.cnr.it/EAGLES96/intro.html

[8] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Contrary to what we obtained in the PAN-AP-13 corpus for Spanish [5], results for gender are better than for age. This is in line with the rest of participants of 2014 which improved more in gender than in age identification with respect to the results obtained in 2013. This was due to the highest number of classes (3 classes in 2013 vs. 5 continuous ones in 2014). Results for blogs and Twitter are better than for social media and reviews. We may explain this because both blogs and Twitter datasets were manually annotated, ensuring that the gender and age of each author is true. On the contrary, in social media and reviews what the authors reported was assumed to be true. Furthermore, in blogs there are enough texts per author in order to obtain a better profile. In fact, although in Twitter each tweet is short (as much 140 characters), there are hundreds of tweets per author. On the other hand, although the quality of social media with respect to the previous year was improved, the data remain with more noise than in blogs. Reviews, where EmoGraphs obtained the worst results, need a special mention. Besides the short texts and the possibility of deceptive information regarding age and gender, reviews are bounded to hotel domain and to the expression of two kinds of emotions: complain or praise.

4.2 The impact of EmoGraphs

We merged the EmoGraph method with the 1,000 more frequent character 6-grams as described previously. In Figure 3 the contribution of EmoGraph to the overall accuracy is presented.

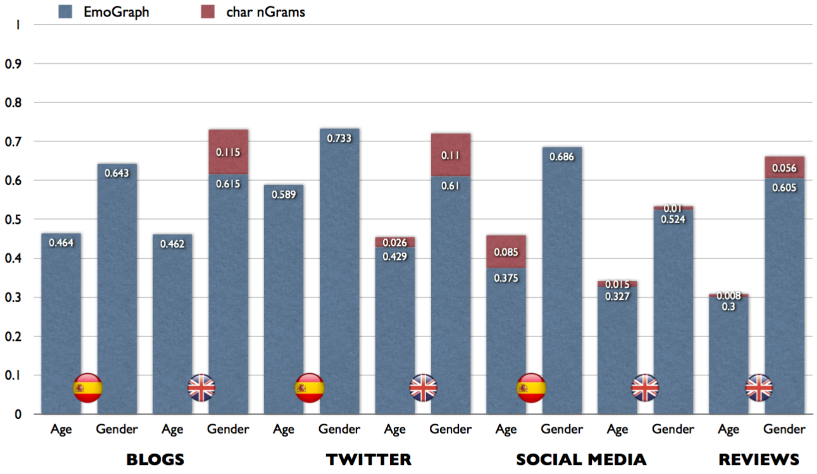

Fig. 3. Contribution to the global accuracy of the EmoGraph representation.

It is noteworthy that, in general, EmoGraph obtains the best results without the need of combination with character n-grams in the Spanish language. This may be due to the lower number of morphosyntactic labels in English. With respect to the performance for gender identification in blogs and Twitter in English, contrary to the one in these social media in Spanish, character n-grams may have helped to capture the missing information due to the lack of detailed morphosyntactic annotation.

5 Conclusions

We have investigated the robustness of the EmoGraph representation for the age and gender identification in several social media genres and different language (Spanish and also English). We showed that our method remains competitive, although it obtains better results in Spanish. In our opinion this is due to the coarse-grained morphosyntactic annotation of English. Results for reviews are very insightful because they are much worse than other genres. We believe this is due to the more bounded topics and emotions.

The performance at age identification in some cases, such as for Spanish blogs, is not much higher than majority class. This is due to the skew in the age distribution.[9] In this sense, it would be interesting to investigated further the application of cost-sensitive machine learning techniques.

References

1. Argamon, S., Koppel, M., Fine, J., Shimoni, A.: Gender,genre,andwritingstyleinformal written texts. In: TEXT, vol. 23, pp. 321-346 (2003)
2. Levin, B.: English verb classes and alternations. University of Chicago Press, Chicago (1993)
3. Mohammad, S.M., Yang, T.: Tracking sentiment in mail: how gender differ on emotional axes. In: In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (2011)
4. Pennebaker, J.W.: The secret life of pronouns: What our words say about us. Bloomsbury Press (2011)
5. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information Processing & Management, Special Issue on Emotion and Sentiment in Social and Expressive Media (In Press) (2015)
6. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In: In: Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org, vol. 1180 (2014)
7. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: In: Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops. CEUR-WS.org, vol. 1179 (2013)
8. Sidorov, G., Miranda-Jimnez, S., Viveros-Jimnez, F., Gelbukh, F., Castro-Snchez, N., Velsquez, F., Daz-Rangel, I., Surez-Guerra, S., Trevio, A., Gordon-Miranda, J.: Empirical study of opinion mining in spanish tweets. In: 11th Mexican International Conference on Artificial Intelligence, MICAI. pp. 1–4 (2012)
9. Strapparava, C., Valitutti, A.: Wordnet-affect: an affective extension of wordnet. In: In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon (2004)

[9] The skew in the distribution is representative of the real use of social media in the different stages of life: `http://jetscram.com/blog/industry-news/social-media-user-statistics-and-age-demographics-2014/`