Author Profile in Social Media: Identifying Information about Gender, Age, Emotions and beyond*

Francisco Rangel
NLE Lab Universitat Politècnica de València / Autoritas Consulting
kico.rangel@gmail.com

Information is a must for social animals that we are the people, the information needs are evolving and the Information Retrieval has to evolve with them. Traditionally, Information Retrieval tried to answer the question "WHAT" with some kind relevance. We try to answer the question "WHO".

Author profiling: gender, age, emotions, personality traits

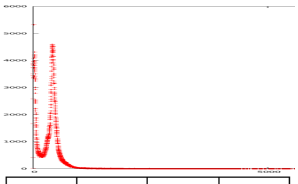
1. INTRODUCTION

Author profiling distinguishes between classes of authors studying their sociolect aspect, that is, how language is shared by people (Koppel et al. 2003) (Argamon et al., 2009). This helps in identifying profiling aspects such as gender, age, native language, or personality type. Author profiling is a problem of growing importance in applications in forensics, security, and marketing. E.g., from a forensic linguistics perspective one would like being able to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence). Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, the demographics of people that like or dislike their products.

2. AGE AND GENDER

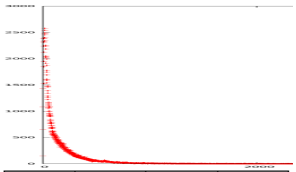
The focus is on author profiling in social media since we are mainly interested in everyday language and how it reflects basic social and personality processes. We have organized a task into PAN 2013[1] for retrieving age and gender from a given text, in Spanish and English. We retrieved a large dataset from blog posts, where each text is labelled with age and gender information.

We retrieved about 30 million of open profiles with the distribution of words per document shown in Figure 1 and Figure 2. As can be seen, there are significant differences between the two languages. More than 80% of Spanish posts are about 15-word long (e.g. greetings, especially for teenagers). On the other hand, English speakers seem also to describe situations, experiences or thoughts, but in a more elaborated way, showing two clear spikes.



MIN	MAX	AVG	STD
0	22,736	335	208

Fig. 1. English distribution



MIN	MAX	AVG	STD
0	12,246	176	832

Fig. 2. Spanish distribution

We grouped posts by author, selecting those authors with at least one post, and chunking in different files those authors with more than 1,000 words in their posts. We also included authors with very few and possibly short posts in order to maintain a realistic evaluation framework. We divided the collection in training, early bird test and final test sets, with the same number for male and female authors. For age detection, we followed what was previously done in (Schler et al., 2006) and considered three classes: 10s (13-17), 20s (23-27) and 30s (33-47). Table 1 shows the corpus statistics with the number of different authors per language, group of age and dataset.

1 Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, Giacomo Inches. http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html

Table 1: *Corpus statistics for training / early bird evaluation/ test datasets*

LANG	AGE	NUM. OF AUTHORS
EN	10s	8,600 / 740 / 888
	20s	42,828 / 3,840 / 4,608
	30s	66,800 / 6,020 / 7,224
ES	10s	1,250 / 120 / 144
	20s	21,300 / 1,920 / 2,304
	30s	15,400 / 1,360 / 1,632

Four teams participated in the early bird evaluation and twenty one in the final one. In Table 2 we show the accuracy of the early bird evaluation and our proposal, based on the features described in Section 3. The best results are bolded.

Table 2: *Early Bird results in terms of accuracy*

TEAM	ENGLISH		SPANISH	
	AGE	GENDER	AGE	GENDER
Gillan	**59.47**	54.13	53.57	47.74
Landra	59.24	**56.31**	57.57	**61.71**
Ayala	2.78	2.77	8.41	8.44
Jankowska	54.63	51.84	44.79	58.34
Baseline	33.24	49.97	33.53	50.01
Rangel	-	-	**62.72**	56.75

It is very important to highlight the difficulty of the task mainly for identifying gender from text, with values similar to the baseline (50%).

3. SIX BASIC EMOTIONS

In (Rangel and Rosso, 2013) we investigate how identifying emotions in Facebook comments in Spanish. We retrieved 1,200 documents from comments made on pages about politics, football and public people, balanced by each theme and by the gender of their authors.

On the basis of stylistic features we aim at identifying the emotional state of the authors in order to profile them. We proposed a SVM based on 60 stylistic features[2] and top 20 words with the highest information gain.

2 Punctuation marks such as dots, commas, quotations, question marks and so on, frequencies such as number of unique words, capital words, words with character flooding and so on, grammatical categories, verb tenses, verb and pronouns number and person, named entities, non-dictionary words, emoticons and emotion words extracted from the Spanish Emotion Lexicon (Sidorov et al., 2012)

We experimented identifying the six basic emotions[3] of Eckman theory, although we do not report results for "*fear*" emotion because too few texts contained words related to this category. Table 3 shows the measures obtained for precision, recall and F.

Table 3: *Identifying emotions from Facebook comments in Spanish language*

EMOTION	PREC.	REC.	F
JOY	71.1	68.6	0.695
ANGER	84.5	73.3	0.772
DISGUST	87.3	75.9	0.799
SURPRISE	67.5	67.8	0.676
SADNESS	91.1	80.2	0.845

The state of the art shows us mainly approaches based on content features, trying to extract the semantics of the sentence. A good summary is offered by (Strapparava & Mihalcea, 2008) who analyses SEMEval-2007[4] task results.

Stylistic features have been used mainly for discovering demographics, such as in PAN task, although some authors also used them in emotion extraction task, as in (Dhaliwal et al., 2007).

Our interest is to link demographics with emotional profile of the user, independently from the content, and stylistic features seem to be key.

We used the same representation model to identify gender, obtaining an accuracy of 53,6 for gender on Facebook comments, and the results shown in Table 2 on the early bird test data set used for the author profiling task at PAN for gender and age.

The results obtained in both tasks, gender and age detection and emotion extraction, suggest us that the stylistic features allow us to detect shared characteristics for demographics and emotional state.

4. BIG FIVE PERSONALITY TRAITS AND BEYOND. FUTURE WORK

Authors like Pennebaker (Pennebaker et al., 2003) connect language use with personality traits, framed into Big Five[5] psychology theory. We aim at going beyond text content, identifying the author's personality in order to predict all demographics from her writing.

3 Joy, anger, disgust, surprise, sadness, fear.

4 http://nlp.cs.swarthmore.edu/semeval/

5 Openness, conscientiousness, extroversion, agreeableness, neuroticism

We focussed our interest on the way the users express themselves, the way they use the language, that is, the style authors write. We conclude that stylistic features help to identify age, gender and emotions of anonymous authors. Our intuition tell us that there is some kind of relation between authors' style of writing and their demographics, emotional profile and, we hope, personality traits. This encourages us to follow the research in this direction in order to understand better how people use language to express themselves and how this could help us to identify the profile of an author.

The features we use for modelling the discursive style are preliminary and simple. As future work we are interested in analysing the discourse in order to investigate further how people use different words of the different grammatical categories, how they place them in the sentence, and how such stylistic decisions provide us information about the author profile.

We must bear in mind with the differences between languages, for example between English and Spanish. For instance, in Spanish the use of pronouns is generally elliptical and it is a choice of the author to use them perhaps to emphasize something, as well as the use of prepositions or determinants in English is more regulated than in Spanish. Due to such specificities, we plan to investigate our proposal to different languages as English.

We also plan to research on the relationship between the demographics such as the gender and age with the emotional profile of the authors and their personality traits, trying to link such tasks in order to build a common framework which allow us to better understand how people use language from a cognitive linguistics viewpoint.

6. REFERENCES

Argamon, S., Koppel, M. Pennebaker, J. & Schler, J. (2009), Automatically profiling the author of an anonymous text, Communications of the ACM 52 (2): 119—123

Dhaliwal, K., Gillies, M., O'connor, J., Oldroyd, A., Robertson, D. & Zhang, L. 2007. Facilitating online role-play using emotionally expressive characters. Artificial and Ambient Intelligence, Proceedings of the AISB Annual Convention, 179-186.

Koppel, M., Argamon, S. & Shimoni, A. (2003). Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), pp. 401-412

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology, (54), 547–577

Rangel, F. & Rosso, P. (2013) Análisis de Emociones en Comentarios de Facebook (submitted)

Schler, J, Koppel, M., Argamon, S. & Pennebaker, J. (2006) Effects of Age and Gender on Blogging. Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., & Gordon, J (2012). Empirical Study of Opinion Mining in Spanish Tweets. LNAI 7629-7630, 14p.

Strapparava, C. & Mihalcea, R. 2008. SemEval-2007 Task 14: Affective Text. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)