Uncovering Plagiarism - Author Profiling at PAN

by Paolo Rosso and Francisco Rangel

PAN is a yearly workshop and evaluation lab on uncovering plagiarism, authorship, and social software misuse.

Since 2009, PAN has been organizing benchmark activities on uncovering plagiarism, authorship, and social software misuse [1]. An additional task - author profiling - has also recently been proposed. Author profiling, instead of focusing on individual authors, studies how language is shared by a class of people. Author profiling is a problem of growing importance in applications in forensics, security and marketing. For instance, a person working in the area of forensic linguistics may need to know the linguistic profile of a suspected text message (language used by a certain type of person) and identify characteristics (with language as evidence). Similarly, from a marketing viewpoint, companies may be interested in determining, through the analysis of blogs and online product reviews, what types of people like or dislike their products.

Author profiling at PAN [2] has been focusing on gender and age identification in social media, both in English and Spanish. We looked for open and public online repositories with posts labelled with author demographics. For age identification, three classes were considered: 10s (13-17), 20s (23-27) and 30s (33-47). We also incorporated a small number of samples from conversations of sexual predators, together with samples from adult-adult sex conversations, with the aim of unveiling fake profiles of potential sexual predators.

With 21 teams, author profiling was one of the most popular tasks at the CLEF conference in 2013. Participants took diverse approaches to the problem: content-based, stylistic-based, n-gram based, etc. Accuracy for gender and age identification, both in English and Spanish, is shown in Figures 1 and 2. Results show the difficulty of the task in a challenging scenario (with 374,100 authors), in particular for gender identification - although the accuracy was slightly higher in Spanish with it being a gender-marked language. With respect to the texts of sexual predators, correct demographics were identified by majority of the participants.

Apart from PAN@CLEF, another two tasks - WCPR@ICWSM and BEA@NAACL-HLT (see links) - were organized in 2013 on predicting different aspects of an author's demographics: specifically, personality traits and native language. This shows the increasing interest of the research community in author profiling. In 2014, PAN will once again be organizing the task on author profiling in social media, as well as tasks on author identification and plagiarism detection.

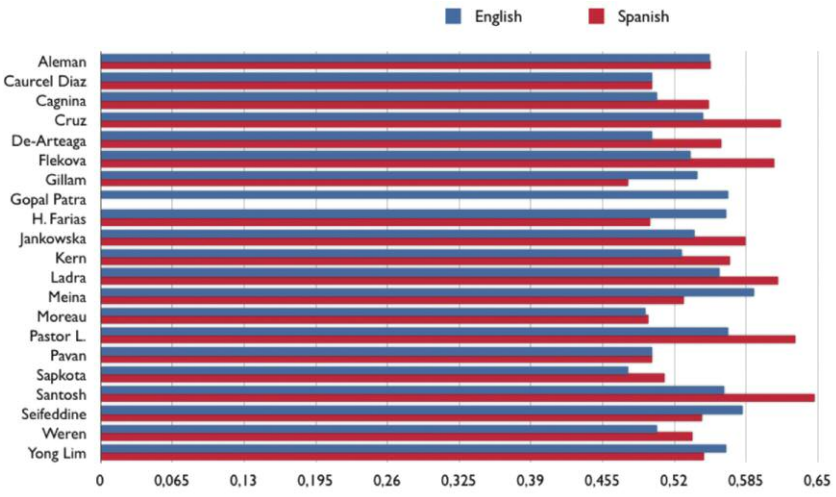

Figure 1: Accuracy for gender identification of social media profiles

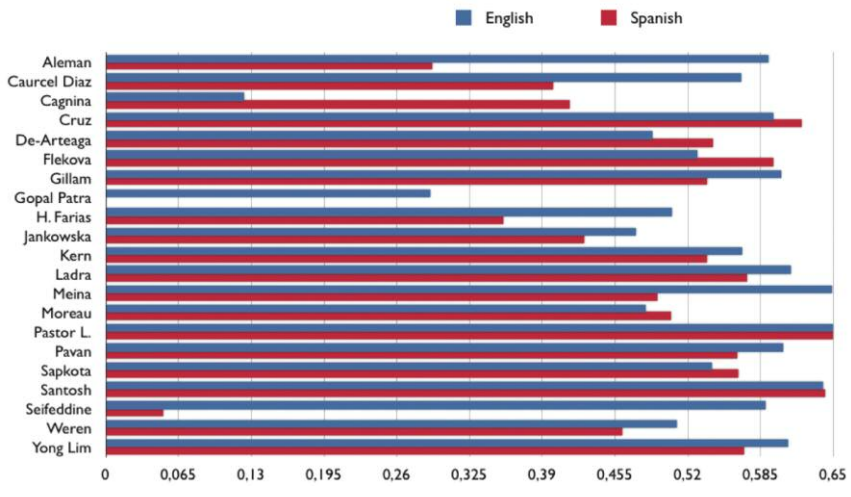

Figure 2: Accuracy for age identification of social media profiles

Links:

PAN: http://pan.webis.de/

http://mypersonality.org/wiki/doku.php?id=wcpr13

http://www.cs.rochester.edu/ ~tetreaul/naacl-bea8.html

References:

[1] T. Gollub, M. Potthast, A. Beyer et al.: "Recent Trends in Digital Text Forensics and its Evaluation: Plagiarism Detection, Author Identification, and Author Profiling", in proc. of the 4th International Conference of the CLEF Initiative (CLEF 13), Springer LNCS 8138, 2013, dx.doi.org/10.1007/978-3-642-40802-1_28

[2] F. Rangel, P. Rosso, M. Koppel et al: "Overview of the Author Profiling Task at PAN 2013. In P. Forner, R. Navigli, D. Tufis, Eds., Working Notes Papers of the CLEF 2013 Evaluation Labs, 2013, http://www.clef-initiative.eu/documents/71612/2e4a4d3a-bae2-47f9-ba3c-552ec66b3e04

Please contact:

Paolo Rosso
Natural Language Engineering Lab, Universitat Politècnica de València, Spain
E-mail: prosso@dsic.upv.es