Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations

Francisco Rangel[1] Paolo Rosso[2] Ben Verhoeven[3]
Walter Daelemans[3] Martin Potthast[4] Benno Stein[4]

[1]Autoritas Consulting, S.A., Spain
[2]PRHLT Research Center, Universitat Politècnica de València, Spain
[3]CLiPS - Computational Linguistics Group, University of Antwerp, Belgium
[4]Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany

pan@webis.de http://pan.webis.de

Abstract This overview presents the framework and the results of the Author Profiling task at PAN 2016. The objective was to predict age and gender from a cross-genre perspective. For this purpose a corpus from Twitter has been provided for training, and different corpora from social media, blogs, essays, and reviews have been provided for evaluation. Altogether, the approaches of 22 participants were evaluated.

1 Introduction

Social media proliferation allows for new communication models and human relations, but often there is a lack of information about who wrote the contents. The possibility of determining people's traits on the basis of what they write is a field of growing interest named author profiling. To infer a user's gender, age, native language or personality traits, simply by analysing her texts, opens a wide range of possibilities from the point of view of forensics, security and marketing. For example, from a forensic viewpoint, to be able to determine the linguistic profile of a person who has written a "suspicious text" may provide valuable background information. Or from a marketing viewpoint, companies may be interested in knowing the demographics of their target audience in order to better segment them.

In the Author Profiling task at PAN 2013 [1] [34], the identification of age and gender relied on a large corpus collected from social media, both in English and Spanish. In PAN 2014[2] [35], we continued focusing on age and gender aspects and, in addition, compiled a corpus of four different genres, namely social media, blogs, Twitter, and hotel reviews. Except for the hotel review subcorpus, which was available in English only, all documents were provided in both English and Spanish. Note that most of the existing research in computational linguistics [3] and social psychology [29] focuses

[1] http://pan.webis.de/clef13/pan13-web/author-profiling.html
[2] http://pan.webis.de/clef14/pan14-web/author-profiling.html

on the English language, and the question is whether the observed relations pertain to other languages and genres as well. In this vein, in PAN 2015[3] [36] we included two new languages, namely Italian and Dutch, along with a new subtask on personality recognition. In PAN 2016[4] [37], we aim at investigating the effect of the cross-genre evaluation: models are trained on one genre, which is Twitter here, and evaluated on another genre different from Twitter.

The paper is organised as follows. Section 2 covers the state of the art, Section 3 describes the corpus and the evaluation measures, and Section 4 presents the approaches submitted by the participants. Section 5 and 6 discuss results and draw conclusions, respectively.

2 Related Work

The study of how certain linguistic features vary according to the profile of their authors is a subject of interest for several different areas such as psychology, linguistics and, more recently, computational linguistics. Pennebaker [30] investigated how the style of writing is associated with personal attributes such as age, gender and personality traits, among others. Argamon *et al.* [3] investigated the task of gender identification on the British National Corpus and achieved approximately 80% accuracy. Similarly in [18] and [9] the authors investigated age and gender identification on formal texts. Recently, most investigations focus on social media: in [20] and [38] the authors investigated the style of writing in blogs. On the other hand, Zhang and Zhang [45] experimented with short segments of blog post and obtained 72.1% accuracy for gender prediction. Similarly, Nguyen *et al.* [27] studied the use of language and age among Dutch Twitter users. Since 2013, a lot of relevant research has been published in the context of the shared task on author profiling organised at PAN [34, 35, 36, 37]. It is worth to mention the second order representation based on relationships between documents and profiles leading to the best results in all editions [22, 23, 2]. Recently, the EmoGraph graph-based approach [33] tried to capture how users convey verbal emotions in the morphosyntactic structure of the discourse, obtaining competitive results with the best performing systems at PAN 2013 and demonstrating its robustness against genres and languages at PAN 2014 [32]. Moreover, the authors in [43] investigated a high variety of different features on the PAN-AP-2013 dataset, and showed the contribution of information retrieval based features in age and gender identification. In [24], the authors approached the task with 3 million features using a MapReduce approach, obtaining high accuracies with a small of processing time.

3 Evaluation Framework

In this section we describe the construction of the corpus and discuss particular properties, challenges, and novelties. Moreover, the evaluation measures are described.

[3] http://pan.webis.de/clef15/pan15-web/author-profiling.html

[4] http://pan.webis.de/clef16/pan16-web/author-profiling.html

3.1 Corpus

In order to study the impact of the cross-genre evaluation on the performance of the different author profiling approaches, we have provided a corpus with different genres for training, early birds and test. The respective subcorpora cover English, Spanish and Dutch. The authors are labelled with age and gender information, except in case of Dutch where only gender information is provided. For labelling age, the following classes were considered: *a*) 18-24; *b*) 25-34; *c*) 35-49; *d*) 50-64; *e*) 65+ .

As in the previous editions, each subcorpus was split into three parts: training, early birds, and test, respectively. The training part was collected from Twitter for the three languages. Early birds and test corpora in the Dutch subcorpus were collected from reviews, whereas in English and Spanish the early birds corpus was collected from social media, and the test corpus was collected from blogs.

Twitter for Training in English and Spanish We have merged the training and test sets from the PAN-AP14 Twitter corpus [10], discarding those Twitter users who deleted or made their account private since then. The corpus was manually selected and annotated in 2014 and is described as follows. Firstly, we looked for public LinkedIn profiles that share a Twitter account. We verified whether the Twitter account exists, whether it is written in one of the languages we are interested in (English or Spanish), whether it is updated only by one person, and whether this person is easily identifiable. We discarded organizational Twitter accounts if we were not sure that the account was updated by the person identified in the LinkedIn profile. Secondly, we looked for age information. Note that in some cases the birth date is published in the user's profile but in most cases it is not, and we looked for the degree starting date in the education section. We used the information shown in Table 1 to define the age range: users whose education were not clear were discarded.

Table 1. Age range mapping using the degree starting date.

Degree starting date	Age group
2006-...	18-24
1997-2006	25-34
1982-1996	35-49
1967-1981	50-64
...-1966	+65

Third, if we could figure out the age, we inferred the gender by the user's photograph and name. Again, for those cases where the gender information was not clear, we discarded the user. The outlined process was done by two independent annotators along with a third person who decided in case of disagreement. Due to Twitter terms of service, we provided the tweets' URLs in order to allow participants to download them. For each Twitter profile, we provided up to 1,000 tweets. The final distribution of the number of authors is shown in the training section of Table 2. The Twitter subcorpus is balanced by gender, i.e., half of the authors are male and the other half are female.

Table 2. Distribution of authors with respect to age classes per language (English and Spanish).

	Training (Twitter)		Early birds (Social Media)		Test (Blogs)	
	English	Spanish	English	Spanish	English	Spanish
18-24	26	16	70	16	10	4
25-34	136	64	92	20	24	12
35-49	182	126	102	16	32	26
50-64	78	38	80	8	10	10
65+	6	6	4	4	2	4
Σ	428	250	348	64	78	56

Social Media for Early Birds in English and Spanish The social media subcorpus was obtained from the test partition of the PAN-AP-14 social media subcorpus, which in turn was obtained by selecting a subset of the PAN-AP-13 corpus. The PAN-AP-14 was build as follows: We have selected authors whose posts have an average number of words greater than 100. We also manually reviewed the documents in order to remove those authors who seemed to have fake profiles such as bots, for example, authors selling the same product (such as mobiles or ads), or authors with a large fraction of text reuse (e.g., teenagers sharing poetry or homework). The final distribution of the number of authors is shown in the early birds columns of Table 2. Again, also the social media subcorpus is balanced by gender.

Blogs for Test in English and Spanish The blog subcorpus for English and Spanish has been obtained from the test partition of the PAN-AP-14 blog subcorpus. The blogs were manually selected and annotated following the methodology described for the Twitter set for Spanish and English. For each blog, we provided up to 25 posts. The final distribution of the number of authors is shown in the test columns of Table 2. The blog subcorpus is balanced by gender as well.

Twitter for Training in Dutch The training data for Dutch are tweets mined as a precursor of TwiSty [41]. TwiSty is a multilingual corpus developed for research in author profiling. It contains personality (MBTI) and gender annotations for a total of 18,168 authors spanning six languages. The Twitter ids of these authors as well as the ids of their available tweets at the time of corpus development are freely available for research purposes. The tweets have undergone language identification and can be found in a Confirmed (as belonging to the language in which the author is situated) and Other category. The final distribution of the Dutch training subcorpus is shown in the first column of Table 3. The Dutch subcorpus is not labelled by age and is balanced by gender, so half of the authors are female and the other half are male.

Table 3. Distribution of authors for Dutch.

Training (Twitter)	Early birds (Reviews)	Test (Reviews)
384	50	500

Reviews for Early birds and Test in Dutch The early birds and test sets for Dutch were reviews from the CSI corpus [40]. The CSI corpus is a yearly expanded corpus of student texts in two genres: essays and reviews. The purpose of this corpus lies primarily in stylometric research, but other applications are possible as well. The available meta-data concerns both the author (gender, age, sexual orientation, region of origin, personality profile) and the document (timestamp, genre, veracity, sentiment, grade). The final distribution of the Dutch early birds and test subcorpus are shown in the second and third columns of Table 3. The early birds set is a random sample of 10% of the test set. The Dutch subcorpus is not labelled by age and is balanced by gender, so half of the authors are female and the other half are male.

3.2 Performance Measures

For evaluating the participants' approaches we have used accuracy. More specifically, we computed the ratio between the number of authors correctly predicted by the total number of authors. We calculated accuracy separately for each language, gender, and age class. Moreover, we obtained the accuracy for the joint identification of age and gender for the languages for which age is available. Then, we averaged the results obtained per language (see Eq. 1).

$$\begin{aligned} \overline{gender} &= \frac{gender\_en + gender\_es + gender\_nl}{3} \\ \overline{age} &= \frac{age\_en + age\_es}{2} \\ \overline{joint} &= \frac{joint\_en + joint\_es}{2} \end{aligned} \tag{1}$$

The final ranking is calculated as the average of the previous values (see Eq. 2):

$$ranking = \frac{\overline{gender} + \overline{age} + \overline{joint}}{3} \tag{2}$$

We computed the statistical significance of performance differences between systems by means of approximate randomisation testing [28].[5] As noted by Yeh [44], for comparing output from classifiers, frequently used statistical significance tests such as paired t-tests make assumptions that do not hold for precision scores and f-scores. Approximate randomisation testing does not make these assumptions and can handle complicated distributions as well as normal distributions. We did a pairwise comparison of accuracies of all systems; given $p < 0.05$, we consider the systems to be significantly different from each other. The complete set of statistical significance tests is illustrated in Appendices A and B for the early birds and test evaluations respectively.

In case of age identification we also measured the average and standard deviation of the distance between the predicted and the true class. We define the distance between classes as the number of hops between them, with the maximum distance equal to 4

[5] We used the implementation by Vincent Van Asch available from the CLiPS website: http://www.clips.uantwerpen.be/scripts/art

in the case of the most distant ones (18-24 and 65+). In the case a participant did not provide a prediction, we added 1 to the maximum distance, penalising this missing value with a distance of 5.

3.3 Software Submissions

We continued to ask software submissions instead of run submissions. With software submissions, participants are asked to submit executables of their author profiling softwares instead of just the output (i.e., runs) of their softwares on a given test set. Our rationale to do so is to increase the sustainability of our shared task and to allow for the re-evaluation of approaches to Author Profiling later on, in particular on future evaluation corpora. To facilitate software submissions, we develop the TIRA experimentation platform [15, 16], which makes handling software submissions at scale as simple as handling run submissions. Using TIRA, participants deploy their software on virtual machines at our site, where we provide a web interface that allows the participants to execute and test their algorithms in a remote fashion [17].

4 Overview of the Submitted Approaches

Twenty two teams participated in the Author Profiling shared task; fourteen of them submitted the notebook paper and one team (*Devalkeneer*) provided us with a description of its approach. We analyse their approaches from three perspectives: pre-processing, features to represent the authors' texts and classification approaches.

Pre-processing Various participants, including *Devalkeneer*, cleaned the HTML and XML to obtain plain text [39, 7, 4]. Lemmatization was applied in [8], although the authors reported no improvement in their results. In [11] the authors applied stemming. In [14, 8, 26] the authors removed punctuation signs, stop words were removed in [11, 1]; the authors in [1, 8] lowercased the texts and digits were removed in [8, 25]. However, the most common pre-processing regarded in Twitter specific components such as hashtags, mentions, RTs or urls [1, 8, 25, 7, 19, 14]. The authors removed or converted Twitter specific elements into constant strings. Some participants applied feature selection such as [14, 4] reporting no effect in the results; the participants in [25] applied transition point techniques.

Features Many participants [42, 4, 8, 7, 14, 26, 31] considered different kinds of stylistic features. For example, the frequency of use of function words, words that are not in a predefined dictionary, slang, capital letters, unique words, and so on so forth. The use of specific sentences per gender (e.g. "my wife", "my man", "my girlfriend"...) and age ("I'm" followed by a number) was used in [14] and sentiment words were taken into account in [31, 14].

Nevertheless, most of the previous participants combined stylistic features with n-grams models [4, 8, 26, 7, 14, 39, 25], parts-of-speech [7, 42, 14, 4], collocations [7], LDA [7], different readability indexes [14], vocabulary richness [4], correctness [31] or verbosity [12]. The second order representation introduced by Pastor *et al.* [22] at PAN'13 has been used by three participants [42, 8, 25].

Devalkeneer modelled the authors with a bag-of-words approach, as well as the authors in [19, 11]. The authors in [1, 12] weighted their n-grams with tf-idf. Finally, this is the first time a participant approached the task with distributed representations (word2vec) [6].

Classification approaches Most of the participants approached the task as a machine learning task. For example, in [4] the authors explored different tree-based algorithms such as Random Forest, as did the participants in [31], besides J48 and LADTree. Logistic regression was used in [26, 7]. In the latter, the authors also used SVM, as did most of the participants [12, 6, 25, 8, 11, 42]. In [14] the authors combined SVM with bootstrapping, and in [1] the authors applied stacking. *Devalkeneer* trained a Class-RBM [21] classifier. Finally, two teams [19, 39] used distance-based approaches to predict the closest class.

5 Evaluation and Discussion of the Submitted Approaches

We divided the evaluation in two steps, providing an early bird option for those participants who wanted to receive a-priori feedback on their performance. There were 17 early bird submissions and eventually 22 for the final evaluation. We show results separately for the evaluation for each corpus part and for each language. The results are given as accuracy values for both the identification of age and gender in isolation and as joint task.

Due to the fact that we provided two different genres for both evaluations for English and Spanish, and also due to the fact that this year's corpus has been obtained from 2014 for English and Spanish, we compare between years and genres in order to deeper investigate the effect of the cross-genre evaluation.

5.1 Early Birds Evaluation

Results for early birds are shown in Table 4 for English and Spanish, and in Table 5 for Dutch. A baseline was provided for comparison purposes, which predicted age and gender randomly. Some participants did not run their systems on not all the languages.

The early birds for English and Spanish were evaluated on a social media corpus as described in Section 3.1. Results for the joint accuracy are very similar for English and Spanish, with the best ones about 20%. With respect to age identification, the best result is obtained for English (38.79%) with about 3% over Spanish (35.94%). Gender identification shows the opposite picture, where the best value for Spanish (70.31%) is much higher than for English (55.75%). In both cases the baseline is amply improved for age and joint identification. However, in case of gender identification for English, most systems obtained accuracies below the baseline.

The best result for age and joint identification for English was obtained by the team named *Waser*. However, the authors withdrew their submission at posteriori so we do not have information about their approach. The best result for gender identification was obtained by Busger et al. [42] who approached the task with combinations of stylistic features such as function words, parts-of-speech, emoticons, punctuations signs along with the second order representation, training their models with SVM. In case of Spanish, three teams obtained the best result for the joint identification. Bougiatiotis and Krithara [8] approached the task by combining stylometric features with character n-grams and the second order representation, training their models with SVM as well. Kocher and Savoy [19] used a distance-based approach and bag-of-words including symbols. Finally, Modaresi et al. [26] used logistic regression with a combination of stylometric and lexical features with word and character n-grams, among others. *Devalkeneer* approached the task using a ClassRBM trained with bag-of-words, obtaining the best result for the age identification, whereas *Waser* obtained the best result for gender identification.

Table 4. Accuracy results for early birds (social media) in terms of accuracy on English (left) and Spanish (right) texts; * = withdrawn.

English			
Team	Joint	Gender	Age
Waser*	**0.2098**	0.5230	**0.3879**
Busger *et al.*	0.1897	**0.5575**	0.3046
Devalkeneer	0.1839	0.5259	0.2931
Dichiu & Rancea	0.1753	0.5345	0.2989
Agrawal & Gonçalves	0.1724	0.5431	0.3103
Bougiatiotis & Krithara	0.1724	0.5345	0.3046
Modaresi(a)	0.1724	0.5057	0.3218
Bilan *et al.*	0.1667	0.5374	0.2902
Gencheva *et al.*	0.1638	0.5287	0.2902
Garciarena *et al.*	0.1609	0.5201	0.2816
Kocher & Savoy	0.1552	0.5144	0.2816
Modaresi *et al.*	0.1552	0.5029	0.3017
Zahid	0.1523	0.4885	0.3103
Ashraf *et al.*	0.1494	0.4971	0.2902
Roman-Gomez	0.1494	0.5144	0.2874
Bakkar *et al.*	0.1466	0.5029	0.2874
baseline	0.1207	0.5402	0.2126
Pimas *et al.*	0.0057	0.0201	0.0086

Spanish			
Team	Joint	Gender	Age
Bougiatiotis & Krithara	**0.2031**	0.5781	0.3438
Kocher & Savoy	**0.2031**	0.5000	0.3125
Modaresi *et al.*	**0.2031**	0.6406	0.2813
Busger *et al.*	0.1875	0.5313	0.2813
Devalkeneer	0.1875	0.5625	**0.3594**
Garciarena *et al.*	0.1875	0.5625	0.2969
Waser*	0.1875	**0.7031**	0.2813
Dichiu & Rancea	0.1719	0.5469	0.2813
Gencheva *et al.*	0.1563	0.6250	0.2656
Bilan *et al.*	0.1406	0.5781	0.2969
Modaresi(a)	0.1406	0.6250	0.2969
Zahid	0.1406	0.5781	0.2969
Agrawal & Gonçalves	0.1094	0.4688	0.2500
Roman-Gomez	0.0938	0.5156	0.1563
baseline	0.0625	0.5313	0.1094

In Appendix A, statistical significances of all pairwise system comparisons are detailed. As can be seen in Table A3, on English *Waser* obtained highly significant results over Busger *et al.* and *Devalkeneer*. Similarly for age identification, *Waser* obtained very significant results over Modaresi *et al.* and significant results over Agrawal and Gonçalves, as shown in Table A2. With respect to gender identification, Busger *et al.* obtained highly significant results over Agrawal and Gonçalves, Bilan *et al.* and Bougiatiotis and Krithara, as shown in Table A1.

With respect to Spanish, the achieved significances are not so clear. For example, although in age identification the best result was obtained by *Devalkeneer*, it is not significant better than Bougiatiotis and Krithara, and Kocher and Savoy, as shown in Table A6. Similarly, in gender identification as can be seen in Table A5, *Waser* obtained very significant results over Gencheva *et al.* but not a significant improvement over Modaresi *et al.* and Modaresi(a). But the main difficulty appears when trying to analyse the significance in joint identification. The best results were obtained in a draw by Bougiatiotis and Krithara, Kocher and Savoy, and Modaresi *et al.* and the second one by Busger *et al.*, *Devalkeneer*, Garciarena *et al.* and *Waser*. For example, Bougiatiotis and Krithara obtained highly significant results over Busger *et al.* whereas no significant results over Garciarena *et al.* A more in-depth analysis can be done looking at Table A7.

For Dutch, results only comprise gender identification. Accuracy values range between 44% and 62%, with half of the participants over the random 50% accuracy and only three of them over the provided baseline. The two best systems obtained higher accuracies (62% and 60%) than the best ones in English (55.75% and 54.31%). In Spanish the best result is about 70.31%, with high difference over the second and third with accuracies of 64,06% and 62,50% respectively. We can see in Table A9 that the best team *Roman-Gomez* obtained no significant improvement over *Waser* but highly significant results over Gencheva *et al.*

Table 5. Accuracy results for early birds on Dutch texts; * = withdrawn.

Team	Gender
Roman-Gomez	**0.6200**
Waser*	0.6000
Gencheva *et al.*	0.5600
baseline	0.5600
Bayot & Gonçalves	0.5400
Bilan *et al.*	0.5400
Dichiu & Rancea	0.5400
Garciarena *et al.*	0.5400
Zahid	0.5400
Kocher & Savoy	0.5200
Agrawal & Gonçalves	0.5000
Busger *et al.*	0.5000
Devalkeneer	0.5000
Modaresi *et al.*	0.5000
Modaresi(a)	0.5000
Poongunran	0.4800
Bougiatiotis & Krithara	0.4400

5.2 Final Evaluation

As for the early birds, we analyse results for English and Spanish together (Table 6) since they share the same genre with both age and gender annotations. In this case the evaluation has been done on blogs (Section 3.1), where we can see higher accuracies for both languages than the early birds evaluated on social media.

On this corpus, results for English and Spanish are more similar. The best accuracy in joint identification is 39.74% in English against 42.86% in Spanish, and similarly in gender identification (75.64% versus 73.21%). However, for age identification, results on English (58.97%) are higher than on Spanish (51.79%). In both languages, most of the participants obtained higher values than the baseline provided.

Busger *et al.* [42] obtained the best result in the age identification for English (58.97%) by using SVM and combinations of stylistic and second order features, whereas the best one for gender identification was obtained by Modaresi *et al.* [26] who used logistic regression with a mix of stylometric, lexical and n-gram features. The best result for the joint identification in English was obtained by Bougiatiotis and Krithara [8] with SVM learnt from stylometric, n-grams and second order features. In Spanish, Modaresi *et al.* obtained the best results in age and joint identification together with Busger *et al.*, whereas *Deneva* obtained the best result in gender identification (73.21%); a description of the system was not provided.

Table 6. Accuracy results (blogs) on English (left) and Spanish (right) texts; * = withdrawn.

English			
Team	Joint	Gender	Age
Bougiatiotis & Krithara	**0.3974**	0.6923	0.5513
Busger *et al.*	0.3846	0.6410	**0.5897**
Modaresi *et al.*	0.3846	**0.7564**	0.5128
Bilan *et al.*	0.3333	0.7436	0.4487
Waser*	0.3205	0.5897	0.4359
Devalkeneer	0.3205	0.6026	0.4487
Modaresi(a)	0.3205	0.6667	0.4487
Markov *et al.*	0.2949	0.6154	0.4487
Roman-Gomez	0.2821	0.6538	0.3974
Dichiu & Rancea	0.2692	0.6154	0.4103
Gencheva *et al.*	0.2564	0.6795	0.3718
Kocher & Savoy	0.2564	0.5769	0.4103
Ashraf *et al.*	0.2564	0.5769	0.3718
Bayot & Gonçalves	0.2179	0.6282	0.3590
Deneva	0.2051	0.5128	0.3718
Bakkar *et al.*	0.2051	0.5385	0.3718
Agrawal & Gonçalves	0.1923	0.5128	0.3846
Zahid	0.1923	0.5000	0.3846
Aceituno	0.1667	0.5000	0.3205
Garciarena *et al.*	0.1538	0.4615	0.3718
Pimas *et al.*	0.1410	0.5769	0.3205
baseline	0.0897	0.5641	0.1923

Spanish			
Team	Joint	Gender	Age
Busger *et al.*	**0.4286**	0.7143	**0.5179**
Modaresi *et al.*	**0.4286**	0.6964	**0.5179**
Bilan *et al.*	0.3750	0.6250	0.4643
Markov *et al.*	0.3750	0.6607	0.4464
Dichiu & Rancea	0.3214	0.6429	0.4643
Bayot & Gonçalves	0.3036	0.5893	0.4821
Modaresi(a)	0.3036	0.6964	0.4464
Devalkeneer	0.2857	0.5179	0.4821
Agrawal & Gonçalves	0.2857	0.5357	0.4821
Deneva	0.2679	**0.7321**	0.3214
Waser*	0.2679	0.5893	0.4107
Bougiatiotis & Krithara	0.2500	0.6786	0.3214
Gencheva *et al.*	0.2500	0.6250	0.3214
Garciarena *et al.*	0.2500	0.5000	0.4286
Zahid	0.2143	0.4821	0.4464
Kocher & Savoy	0.1964	0.5357	0.3393
Roman-Gomez	0.1250	0.5000	0.2500
baseline	0.1250	0.5000	0.1786
Aceituno	0.0893	0.4643	0.2143

In Appendix B, statistical significances of all pairwise systems comparisons are detailed. In English, it is shown in Tables B1 and B2 that the three first systems are not significantly different for age and gender identification. Similarly with respect to the joint identification, Bougiatiotis and Kithara obtained highly significant results over Busger *et al.* although no significant results over Modaresi *et al.* In Spanish, something similar occurs with joint identification, as show in Table B7, where Busger *et al.* and Modaresi *et al.* shared the first position but without significance with Bilan *et al.* and Markov *et al.* that shared the second position. With respect to gender identification, *Deneva* obtained no significant results over Busger *et al.*, although highly significant over Modaresi *et al.*, as shown in Table B5. However, the comparison in age identification is more complex. Busger *et al.* and Modaresi *et al.* shared the first position, but only Busger *et al.* obtained highly significant results over Bayot and Gonçalves whereas Modaresi *et al.* did it with *Devalkeneer*, not having significance in the other way around. With respect to Agrawal and Gonçalves, Busger *et al.* obtained highly significant results whereas Modaresi *et al.* did not obtain significant improvements.

In Table 7 results for gender identification for Dutch are shown. We can see that the accuracies are much lower than for English and Spanish. The highest accuracy for Dutch is 61.80% versus 75.64% and 73.21% respectively for English and Spanish. As expected, these results are more similar to the ones obtained for the early birds (evaluated for social media) for English, where the best participant obtained 55.75%. Bayot and Gonçalves [6] obtained the best result by training a SVM with word2vec features.

We can see in Table B9 that the best performing team Bayot and Gonçalves obtained no significant improvement over *Roman-Gomez* and Bilan *et al.*

Table 7. Accuracy results on Dutch texts; * = withdrawn.

Team	Gender
Deneva	**0.6180**
Bayot & Gonçalves	0.5680
Roman-Gomez	0.5620
Bilan *et al.*	0.5500
Waser*	0.5320
baseline	0.5300
Dichiu & Rancea	0.5260
Garciarena *et al.*	0.5260
Poongunran	0.5140
Gencheva *et al.*	0.5100
Markov *et al.*	0.5100
Agrawal & Gonçalves	0.5080
Devalkeneer	0.5060
Aceituno	0.5040
Kocher & Savoy	0.5040
Modaresi *et al.*	0.5040
Busger *et al.*	0.5000
Modaresi(a)	0.5000
Bougiatiotis & Krithara	0.4160

5.3 Comparison Between Genres

In this section we compare results obtained for both genres (social media and blogs) and show them in Tables 8 and 9 for English and Spanish respectively. We only show results for those participants who evaluated their systems with both datasets. Teams are alphabetically ordered and the best results are highlighted in bold. As can be seen, in general there is an improvement on the results from social media to blogs in both languages. We highlighted in italics those results that are lower in blogs than in social media.

The best results for English are 58.97%, 75.64% and 39.74% for age, gender and joint identification respectively against 38.79%, 55.75% and 20.98% obtained in social media. They involve an increase of about 52%, 36% and 89% respectively. If we analyse the descriptive statistics provided in the bottom side of the table, we can see an increase of all of them. Concretely, the average results increase 74%, 24% and 54% from social media to blogs. We can conclude that the performance in blogs is significantly higher than in social media, which means that the cross-genre effect for social media is larger.

Table 8. Accuracy results contrasting early birds (social media) with test (blogs) on English; * = withdrawn.

	Joint		**Gender**		**Age**	
Team	Social Media	Blogs	Social Media	Blogs	Social Media	Blogs
Agrawal & Gonçalves	0.1724	0.1923	0.5431	*0.5128*	0.3103	0.3846
Ashraf *et al.*	0.1494	0.2564	0.4971	0.5769	0.2902	0.3718
Bakkar *et al.*	0.1466	0.2051	0.5029	0.5385	0.2874	0.3718
Bilan *et al.*	0.1667	0.3333	0.5374	0.7436	0.2902	0.4487
Bougiatiotis & Krithara	0.1724	**0.3974**	0.5345	0.6923	0.3046	0.5513
Busger *et al.*	0.1897	0.3846	**0.5575**	0.6410	0.3046	**0.5897**
Devalkeneer	0.1839	0.3205	0.5259	0.6026	0.2931	0.4487
Dichiu & Rancea	0.1753	0.2692	0.5345	0.6154	0.2989	0.4103
Garciarena *et al.*	0.1609	*0.1538*	0.5201	*0.4615*	0.2816	0.3718
Gencheva *et al.*	0.1638	0.2564	0.5287	0.6795	0.2902	0.3718
Kocher & Savoy	0.1552	0.2564	0.5144	0.5769	0.2816	0.4103
Modaresi(a)	0.1724	0.3205	0.5057	0.6667	0.3218	0.4487
Modaresi *et al.*	0.1552	0.3846	0.5029	**0.7564**	0.3017	0.5128
Pimas *et al.*	0.0057	0.1410	0.0201	0.5769	0.0086	0.3205
Roman-Gomez	0.1494	0.2821	0.5144	0.6538	0.2874	0.3974
Waser*	**0.2098**	0.3205	0.5230	0.5897	**0.3879**	0.4359
Zahid	0.1523	0.1923	0.4885	0.5000	0.3103	0.3846
Min	0.0057	0.1410	0.0201	0.4615	0.0086	0.3205
Q1	0.1523	0.2051	0.5029	0.5769	0.2874	0.3718
Median	0.1638	0.2692	0.5201	0.6026	0.2931	0.4103
Mean	0.1577	0.2745	0.4912	0.6109	0.2853	0.4253
SDev	0.0425	0.0794	0.1227	0.0827	0.0754	0.0704
Q3	0.1724	0.3205	0.5345	0.6667	0.3046	0.4487
Max	0.2098	0.3974	0.5575	0.7564	0.3879	0.5897

Similarly, in Spanish the best results obtained in blogs are 51.79%, 71.43% and 42.86% against 35.94%, 70.31% and 20.31% in social media, respectively for age, gender and joint identification. They imply an improvement of about 44%, 2% and 111% respectively. We can see that the increment in the gender identification is subtle. If we analyse the descriptive statistics provided in the bottom side of the table, we can see an increase of all of them, except for the first quartile in gender identification. Concretely, the average results increase 72%, 4% and 47% from social media to blogs. We can conclude that also for Spanish the performance in blogs is significantly higher than in social media, except in gender identification where results are quite similar. This means that the cross-genre effect affected in a greater way to social media, except as it was said, in case of gender identification.

Table 9. Accuracy results contrasting early birds (social media) with test (blogs) on Spanish; * = withdrawn.

	Joint		**Gender**		**Age**	
Team	Social Media	Blogs	Social Media	Blogs	Social Media	Blogs
Agrawal & Gonçalves	0.1094	0.2857	0.4688	0.5357	0.2500	0.4821
Bilan *et al.*	0.1406	0.3750	0.5781	0.6250	0.2969	0.4643
Bougiatiotis & Krithara	**0.2031**	0.2500	0.5781	0.6786	0.3438	*0.3214*
Busger *et al.*	0.1875	**0.4286**	0.5313	**0.7143**	0.2813	**0.5179**
Devalkeneer	0.1875	0.2857	0.5625	*0.5179*	**0.3594**	0.4821
Dichiu & Rancea	0.1719	0.3214	0.5469	0.6429	0.2813	0.4643
Garciarena *et al.*	0.1875	0.2500	0.5625	*0.5000*	0.2969	0.4286
Gencheva *et al.*	0.1563	0.2500	0.6250	0.6250	0.2656	0.3214
Kocher & Savoy	**0.2031**	*0.1964*	0.5000	0.5357	0.3125	0.3393
Modaresi(a)	0.1406	0.3036	0.6250	0.6964	0.2969	0.4464
Modaresi *et al.*	**0.2031**	**0.4286**	0.6406	0.6964	0.2813	**0.5179**
Roman-Gomez	0.0938	0.1250	0.5156	*0.5000*	0.1563	0.2500
Waser*	0.1875	0.2679	**0.7031**	*0.5893*	0.2813	0.4107
Zahid	0.1406	0.2143	0.5781	*0.4821*	0.2969	0.4464
Min	0.0938	0.1250	0.4688	0.4821	0.1563	0.2500
Q1	0.1406	0.2500	0.5352	*0.5224*	0.2813	0.3572
Median	0.1797	0.2768	0.5703	0.6072	0.2891	0.4464
Mean	0.1652	0.2844	0.5725	0.5957	0.2857	0.4209
SDev	0.0356	0.0848	0.0615	0.0831	0.0468	0.0819
Q3	0.1875	0.3170	0.6133	0.6697	0.2969	0.4776
Max	0.2031	0.4286	0.7031	0.7143	0.3594	0.5179

Differently to gender identification, which consisted of a binary classification, in the case of age identification there are five different possibilities. Therefore, the accuracy measure does not give a complete picture of the situation. In such a case, we aimed at investigating the distance between the predicted classes and the truth ones as described in Section 3.2, and how the cross-genre evaluation may affect to the obtained values. In Figure 1 and Table 10 the average and standard deviation of the distances between predicted and true classes per subcorpus are shown. The highest distances on average are produced for social media, especially in Spanish with a value of 1.0379 ± 0.4289. The lowest distances on average are obtained for blogs in both languages. The raise of distance from social media to blogs in the corresponding languages means that age identification in blogs is less affected by the cross-genre evaluation than in social media.

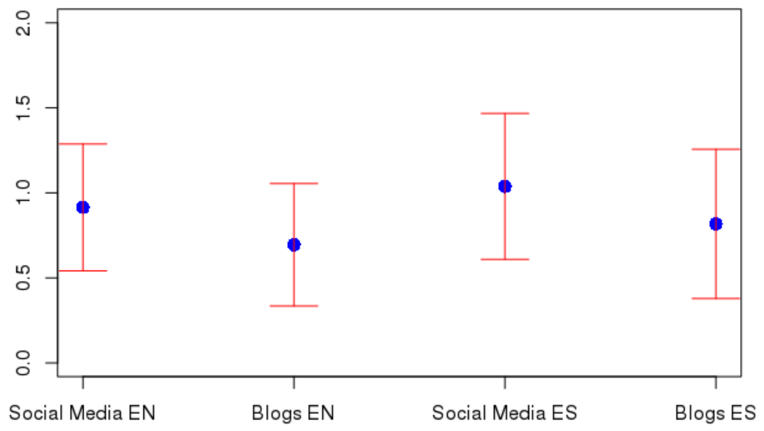

Figure 1. Distances between predicted classes and true classes per subcorpus and language.

Table 10. Distances between predicted classes and true classes per subcorpus and language.

	English		**Spanish**	
	Social Media	Blogs	Social Media	Blogs
Mean	0.9146	0.6951	1.0379	0.8176
SDev	0.7457	0.7199	0.8579	0.8775

5.4 Comparison between Years

It is noteworthy that for English and Spanish the evaluation was carried out with the same social media and blog partitions than in PAN'14. Taking into account social media, in 2014 the best results for English were 20.62%, 54.21% and 36.52% for joint, gender and age identification respectively. These values are very similar to the best results obtained in 2016: 20.98%, 55.75% and 38.79%. In Figures 2 and 3 the corresponding distributions are shown. We can see that both distributions for 2014 and 2016 are quite similar, maybe with more sparsity in 2014 for age and joint identification. In these subtasks, results are higher in 2014 except for one participant (*Waser*) that outperformed the rest in both years. In gender identification results are more similar, maybe with more sparsity in the case of 2016 and also with more participants with higher results. We can conclude that in this case the cross-genre had no effect on the evaluation, although this may be due to the fact that in both years results were very close to the random baseline.

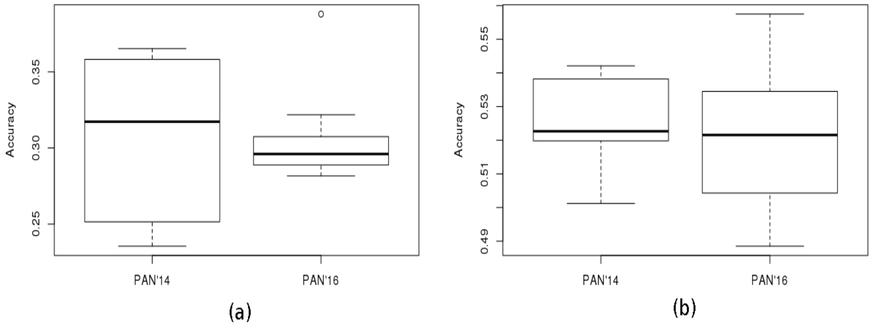

Figure 2. Distribution of results for English Social Media for (a) age identification and (b) gender identification.

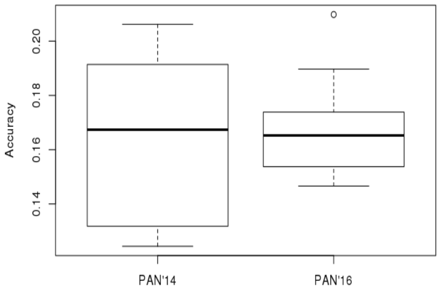

Figure 3. Distribution of results for English Social Media for joint identification.

For Spanish, the corresponding distributions are shown in Figures 4 and 5. For gender identification results seem to be quite similar in the two editions, also with similar results for the best ones (68.37% in 2014 vs. 70.31% in 2016). For joint identification, results falls from 33.57% to 20.31%, and from 48.94% to 35.94% for age identification. In both cases the corresponding distributions show a strong descent in the performance. Hence, in case of Spanish we can conclude that the cross-genre evaluation in social media of models trained on Twitter, had a strong impact on the joint and age identification. However, it seems that the gender identification is not affected too much.

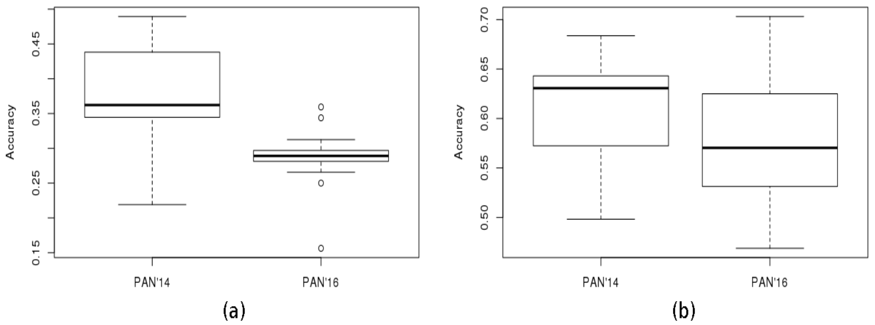

Figure 4. Distribution of results for Spanish Social Media for (a) age identification and (b) gender identification.

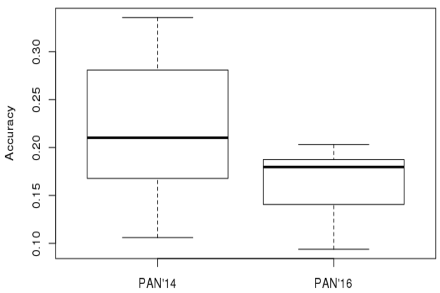

Figure 5. Distribution of results for Spanish Social Media for joint identification.

In the case of blogs, the distribution of results is shown in Figures 6 and 7 for English and Figures 8 and 9 for Spanish. As can be seen, results obtained in 2016 are higher than those obtained in 2014 for all the subtasks and languages. We should highlight the higher results in age identification in English, where most of the systems obtained higher values than the best performing ones in 2014, and also the deviation among teams is smaller.

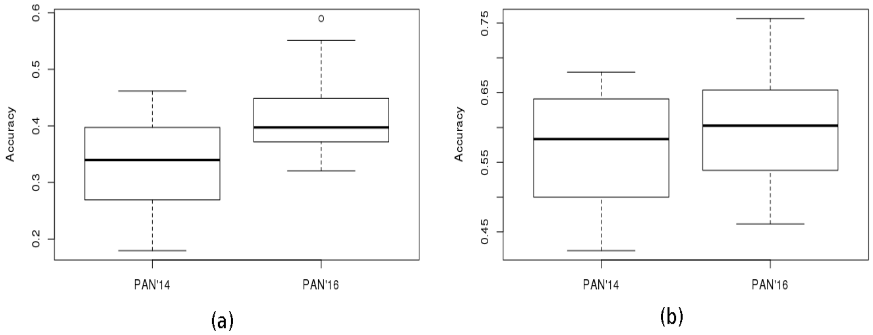

Figure 6. Distribution of results for English blogs for (a) age identification and (b) gender identification.

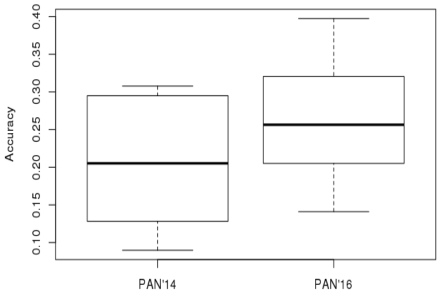

Figure 7. Distribution of results for English blogs for joint identification.

As reported earlier, results for Spanish are better than in 2014, especially for gender and joint identification, where most of the systems obtained higher results than the best ones in 2014. Results on age identification seem very similar in both years.

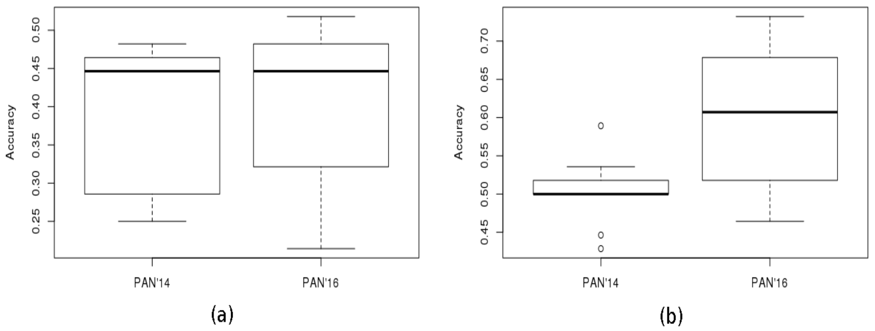

Figure 8. Distribution of results for Spanish blogs for (a) age identification and (b) gender identification.

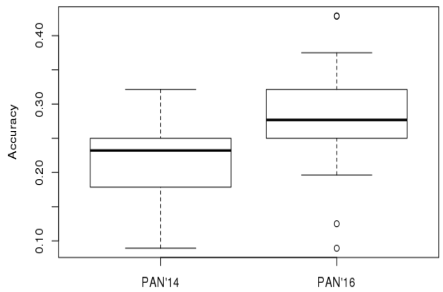

Figure 9. Distribution of results for Spanish blogs for joint identification.

In Tables 11 and 12 comparative statistics among genres are shown. Concretely, we have compared the distribution of results obtained in 2014 on Twitter with this year results on social media and blogs[6]. In case of English we can see that the results obtained on Twitter were higher that in social media for all the statistics. For example, the mean of the results raised from 15.77%, 49.12% and 28.53% to 22.65%, 58.66% and 37.81% in joint, gender and age respectively. Something similar occurred with the best results that raised from 20.98%, 55.75% and 38.79% to 35.71%, 73.38% and 50.65% respectively for joint, gender and age predictions. However, results on blogs were higher than on Twitter. We can observe an increase on average from 22.65%, 58.66% and

[6] It should be taken into account that the results are not completely comparable since the participants in both editions are not the same, but we can obtain some insights about the effect of the cross-genre evaluation depending on the evaluation genre.

37.81% obtained on Twitter to 27.45%, 61.09% and 42.53% on blogs respectively for joint, age and gender identification. Similarly, there was an increment in the highest results from 35.71%, 73.38% and 50.65% on Twitter to 39.74%, 75.64% and 58.97% on blogs, respectively for joint, gender and age identification.

Table 11. Comparative statistics among genres in English. Twitter represents the evaluation in the same genre carried out in 2014. Social media (SM) and blogs represent the cross-genre evaluation.

	Joint			**Gender**			**Age**		
Team	Twitter	SM	Blogs	Twiter	SM	Blogs	Twitter	SM	Blogs
Min	0.0584	0.0057	0.1410	0.5000	0.0201	0.4615	0.1104	0.0086	0.3205
Q1	0.1948	0.1523	0.2051	0.5130	0.5029	0.5769	0.3377	0.2874	0.3718
Median	0.2013	0.1638	0.2692	0.5390	0.5201	0.6026	0.3896	0.2931	0.4103
Mean	0.2265	0.1577	0.2745	0.5866	0.4912	0.6109	0.3781	0.2853	0.4253
SDev	0.0957	0.0425	0.0794	0.0948	0.1227	0.0827	0.1177	0.0754	0.0704
Q3	0.3052	0.1724	0.3205	0.6688	0.5345	0.6667	0.4416	0.3046	0.4487
Max	0.3571	0.2098	0.3974	0.7338	0.5575	0.7564	0.5065	0.3879	0.5897

With respect to Spanish, the results were lower on social media than on Twitter, except for gender identification where the mean was similar for both genres (57.36% on Twitter vs. 57.25% on social media). Nevertheless, the maximum value was achieved on social media (70.31%) over Twitter (65.56%). With respect to blogs, in some cases results obtained on Twitter were higher. For example, in the joint identification the mean on Twitter (28.89%) was slightly higher than on blogs (28.44%) as well as the maximum value (43.33% on Twitter vs. 42.86% on blogs). Similarly, in age identification the mean and the maximum values were higher on Twitter (48.75% and 61.11%) than on blogs (42.09% and 51.79%). On the contrary, results for gender identification were higher on blogs than on Twitter, with an increase in the mean and the maximum values from 57.36% to 59.57% and from 65.56% to 71.43% respectively.

Table 12. Comparative statistics among genres in Spanish. Twitter represents the evaluation in the same genre carried out in 2014. Social media (SM) and blogs represent the cross-genre evaluation.

	Joint			**Gender**			**Age**		
Team	Twitter	SM	Blogs	Twiter	SM	Blogs	Twitter	SM	Blogs
Min	0.1444	0.0938	0.1250	0.5000	0.4688	0.4821	0.2222	0.1563	0.2500
Q1	0.2528	0.1406	0.2500	0.5277	0.5352	0.5224	0.4972	0.2813	0.3572
Median	0.2944	0.1797	0.2768	0.5722	0.5703	0.6072	0.5111	0.2891	0.4464
Mean	0.2889	0.1652	0.2844	0.5736	0.5725	0.5957	0.4875	0.2857	0.4209
SDev	0.0871	0.0356	0.0848	0.0588	0.0615	0.0831	0.1137	0.0468	0.0819
Q3	0.3278	0.1875	0.3170	0.6166	0.6133	0.6697	0.5250	0.2969	0.4776
Max	0.4333	0.2031	0.4286	0.6556	0.7031	0.7143	0.6111	0.3594	0.5179

Comparing results obtained on social media and blogs, both in 2014 and 2016, we can conclude that there was no cross-genre effect on social media data, especially in age and joint identification in English and Spanish, with a lightweight effect in gender identification. On the other hand, cross-genre evaluation seems to have an impact when evaluating with blogs, especially in age and joint identification in English and gender and joint identification in Spanish.

Comparing results obtained in Twitter in 2014 (mono-genre) with results obtained on social media and blogs in 2016 (cross-genre), we can conclude that for English there was a cross-genre effect on social media whereas the contrary happened on blogs, where results were higher than on Twitter. In Spanish, the cross-genre affected especially in the age and joint identification whereas it favoured the gender identification.

5.5 Final ranking

In Table 13 the overall performance per language and users' ranking are shown. We can observe that in general, accuracies in both English and Spanish datasets are similar, although the highest results were achieved in Spanish (42.86%). With respect to Dutch, were only the gender accuracy is shown, results are not much better than the random baseline (the highest value is equal to 61.80%).

Table 13. Global ranking by averaging joint accuracy per language; * = withdrawn.

Ranking	Team	Global	English	Spanish	Dutch
1	Busger *et al.*	**0.5258**	0.3846	**0.4286**	0.4960
2	Modaresi *et al.*	0.5247	0.3846	**0.4286**	0.5040
3	Bilan *et al.*	0.4834	0.3333	0.3750	0.5500
4	Modaresi(a)	0.4602	0.3205	0.3036	0.5000
5	Markov *et al.*	0.4593	0.2949	0.3750	0.5100
6	Bougiatiotis & Krithara	0.4519	**0.3974**	0.2500	0.4160
7	Dichiu & Rancea	0.4425	0.2692	0.3214	0.5260
8	Devalkeneer	0.4369	0.3205	0.2857	0.5060
9	Waser*	0.4293	0.3205	0.2679	0.5320
10	Bayot & Gonçalves	0.4255	0.2179	0.3036	0.5680
11	Gencheva *et al.*	0.4015	0.2564	0.2500	0.5100
12	Deneva	0.4014	0.2051	0.2679	**0.6180**
13	Agrawal & Gonçalves	0.3971	0.1923	0.2857	0.5080
14	Kocher & Savoy	0.3800	0.2564	0.1964	0.5040
15	Roman-Gomez	0.3664	0.2821	0.1250	0.5620
16	Garciarena *et al.*	0.3660	0.1538	0.2500	0.5260
17	Zahid	0.3154	0.1923	0.2143	-
18	Aceituno	0.2949	0.1667	0.0893	0.5040
19	Ashraf *et al.*	0.1688	0.2564	-	-
20	Bakkar *et al.*	0.1560	0.2051	-	-
21	Pimas *et al.*	0.1410	0.1410	-	-
22	Poonguran	0.0571	-	-	0.5140

In Table 14 the best results per language and task are shown. We can observe that results for gender identification in English and Spanish are quite similar, and much higher than in Dutch.

Table 14. Best results per language and task for the test set.

	Age and Gender		
Language	*Joint*	Gender	Age
English	0.3974	0.7564	0.5897
Spanish	0.4286	0.7321	0.5179
Dutch	-	0.6180	-

6 Conclusion

In this paper we presented the results of the 4th International Author Profiling Shared Task at PAN-2016 within CLEF-2016. Given Twitter authors for training, the 22 participants had to identify age and gender in a cross-genre framework in English, Spanish and Dutch.

The participants used several different feature types to approach the problem: content-based (bag of words, word n-grams, term vectors, named entities, dictionary words, slang words, contractions, sentiment words, and so on) and stylistic-based features (frequencies, punctuations, POS, Twitter specific elements, readability measures, and so forth). Distributed representations were for the first time used, and several participants used the second order representation that obtained the best performance in the previous three editions. It is difficult to highlight the contribution of any particular feature since the participants used many of them. The second order representation was used by teams that achieved first positions in some of the tasks. Likewise, the distributed representations achieved the first position in gender identification on the Dutch final evaluation.

The early birds evaluation showed higher accuracies in gender identification in Spanish than in English, where most of the participants obtained results below the baseline, such as in Dutch. In the final evaluation, results were similar for English and Spanish. In both cases most of the participants obtained results over the baseline. On the contrary, results in Dutch were significantly weaker, with most participants below the baseline.

Due to the fact that for English and Spanish we provided different genres for early birds (social media) and final test (blogs), a comparison between them provides us with some insights. In both languages, results on blogs were higher than on social media, except in the case of gender identification in Spanish. Similarly when analysing the distances between predicted ages and true ones, they decreased on average from social media to blogs in both languages. Both analyses may suggest a higher effect of the cross-genre set-up on social media than on blogs.

As English and Spanish datasets were based on the PAN'14 ones, the comparison between years allows to draw some conclusions. There is no strong effect of the cross-genre evaluation on English social media, although this may be due to the low results obtained both years on this genre. With respect to Spanish social media, there is a strong impact on joint and age identification, although the gender identification is not affected too much. In blogs the cross-genre effect is positive, especially on age and joint identification in English and gender and joint identification in Spanish. The previous conclusions suggest that – depending on the combination of genres – the cross-genre learning may improve the final result. For example, learning with Twitter where people share their comments without censorship, in a spontaneous way, and where researchers can obtain a high number of texts per author, could be a a good manner to improve the performance of author profiling tasks in other genres (such as blogs) for which it is more difficult to obtain sufficient training data.

Acknowledgements Our special thanks go to all of PAN's participants, as well as to MeaningCloud[7] for sponsoring the author profiling shared task award, and to Adobe[8] for sponsoring PAN. The work of the first author was partially supported by Autoritas Consulting and by Ministerio de Economía y Competitividad de España under grant ECOPORTUNITY IPT-2012-1220-430000. The work of the second author was partially supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030).

Bibliography

1. Madhulika Agrawal and Teresa Gonçalves. Age and gender identification using stacking for classification. In Balog et al. [5].
2. Miguel-Angel Álvarez-Carmona, A.-Pastor López-Monroy, Manuel Montes-Y-Gómez, Luis Villaseñor-Pineda, and Hugo Jair-Escalante. Inaoe's participation at pan'15: author profiling task—notebook for pan at clef 2015. 2015.
3. Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346, 2003.
4. Shaina Ashraf, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. Cross-genre author profile prediction using stylometry-based approach. In Balog et al. [5].
5. Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors. *CLEF 2016 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1609/*, 2016.
6. Roy Khristopher Bayot and Teresa Gonçalves. Author profiling using svms and word embedding averages. In Balog et al. [5].
7. Ivan Bilan and Desislava Zhekova. Caps: A cross-genre author profiling system - notebook for pan at clef 2016. In Balog et al. [5].
8. Konstantinos Bougiatiotis and Anastasia Krithara. Author profiling using complementary second order attributes and stylometric features. In Balog et al. [5].
9. John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[7] http://www.meaningcloud.com/

[8] http://www.adobe.com/

10. Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors. *CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/ Vol-1180/*, 2014.
11. Rodwan Bakkar Deyab, José Duarte, and Teresa Gonçalves. Author profiling using support vector machines. In Balog et al. [5].
12. Daniel Dichiu and Irina Rancea. Using machine learning algorithms for author profiling in social media. In Balog et al. [5].
13. Pamela Forner, Roberto Navigli, and Dan Tufis, editors. *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, 2013.
14. Pepa Gencheva, Martin Boyanov, Elena Deneva, Preslav Nakov, Georgi Georgiev, Yasen Kiprov, and Ivan Koychev. Pancakes team: a composite system of domain-agnostic features for author profiling. In Balog et al. [5].
15. Tim Gollub, Benno Stein, and Steven Burrows. Ousting ivory tower research: towards a web framework for providing experiments as a service. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 1125–1126. ACM, August 2012. ISBN 978-1-4503-1472-5.
16. Tim Gollub, Benno Stein, Steven Burrows, and Dennis Hoppe. TIRA: Configuring, executing, and disseminating information retrieval experiments. In A Min Tjoa, Stephen Liddle, Klaus-Dieter Schewe, and Xiaofang Zhou, editors, *9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA*, pages 151–155, Los Alamitos, California, September 2012. IEEE. ISBN 978-1-4673-2621-6.
17. Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Recent trends in digital text forensics and its evaluation. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13)*, pages 282–302, Berlin Heidelberg New York, September 2013. Springer. ISBN 978-3-642-40801-4.
18. Janet Holmes and Miriam Meyerhoff. *The handbook of language and gender*. Blackwell Handbooks in Linguistics. Wiley, 2003.
19. Mirco Kocher and Jacques Savoy. Unine at clef 2016: author profiling. In Balog et al. [5].
20. Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. literary and linguistic computing 17(4), 2002.
21. H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *25th International Conference on Machine Learning pp. 536–543. ICML'08, ACM*, 2008.
22. A. Pastor Lopez-Monroy, Manuel Montes-Y-Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esau Villatoro-Tello. INAOE's participation at PAN'13: author profiling task—Notebook for PAN at CLEF 2013. In Forner et al. [13].
23. A. Pastor López-Monroy, Manuel Montes y Gómez, Hugo Jair-Escalante, and Luis Villase nor Pineda. Using intra-profile information for author profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [10].

24. Suraj Maharjan, Prasha Shrestha, Thamar Solorio, and Ragib Hasan. A straightforward author profiling approach in mapreduce. In *Advances in Artificial Intelligence. Iberamia*, pages 95–107, 2014.
25. Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, and Alexander Gelbukh. Adapting cross-genre author profiling to language and corpus. In Balog et al. [5].
26. Pashutan Modaresi, Matthias Liebeck, and Stefan Conrad. Exploring the effects of cross-genre machine learning for author profiling in pan 2016. In Balog et al. [5].
27. Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "how old do you think i am?"; a study of language and age in twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.
28. Eric W. Noreen. *Computer intensive methods for testing hypotheses: an introduction*. Wiley, New York, 1989.
29. James W. Pennebaker. *The secret life of pronouns: what our words say about us*. Bloomsbury USA, 2013.
30. James W. Pennebaker, Mathias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.
31. Oliver Pimas, Andi Rexha, Mark Kroll, and Roman Kern. Profiling microblog authors using concreteness and sentiment - know-center at pan 2016 author profiling. In Balog et al. [5].
32. Francisco Rangel and Paolo Rosso. On the multilingual and genre robustness of emographs for author profiling in social media. In *6th international conference of CLEF on experimental IR meets multilinguality, multimodality, and interaction*, pages 274–280. Springer-Verlag, LNCS(9283), 2015.
33. Francisco Rangel and Paolo Rosso. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92, 2016.
34. Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *Forner P., Navigli R., Tufis D. (Eds.), CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org, vol. 1179*, 2013.
35. Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org, vol. 1180*, 2014.
36. Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd author profiling task at pan 2015. In *Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391*, 2015.
37. Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2016.
38. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205. AAAI, 2006.

39. Ma. José Garciarena Ucelay, Ma. Paula Villegas, Dario G. Funez, Leticia C. Cagnina, Marcelo L. Errecalde, Gabriela Ramírez-De-La-Rosa, and Esau Villatoro-Tello. Profile-based approach for age and gender identification. notebook for pan at clef 2016. In Balog et al. [5].
40. Ben Verhoeven and Walter Daelemans. Clips stylometry investigation (csi) corpus: a dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.
41. Ben Verhoeven, Walter Daelemans, and B. Plank. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.
42. Mart Busger Op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. Gronup: Groningen user profiling. In Balog et al. [5].
43. Edson Weren, Anderson Kauer, Lucas Mizusaki, Viviane Moreira, Palazzo de Oliveira, and Leandro Wives. Examining multiple features for author profiling. In *Journal of Information and Data Management*, pages 266–279, 2014.
44. Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, pages 947–953, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
45. Cathy Zhang and Pengyu Zhang. Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA, 2010.

Appendix A Pairwise Comparison of all Systems in social media

For all subsequent tables, the significance levels are encoded as follows:

Symbol	Significance Level		
-		∼	not evaluated
=	p > 0.05	∼	not significant
*	0.05 ≥ p > 0.01	∼	significant
**	0.01 ≥ p > 0.001	∼	very significant
***	p ≤ 0.001	∼	highly significant

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		***	=	-	=	=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Ashraf			***	-	***	***	=	=	***	***	=	=	-	***	***	***	-	***	***	=
Bakkar				-	=	=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Bayot					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bilan						=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Bougiatiotis							***	***	=	=	***	***	-	=	=	=	-	=	=	***
Busger								=	***	***	=	=	-	***	***	***	-	***	***	=
Devalkeneer									***	***	=	=	-	***	***	***	-	***	***	=
Dichiu										=	***	***	-	=	=	=	-	=	=	***
Garciarena											***	***	-	=	=	=	-	=	=	***
Gencheva												=	-	***	***	***	-	***	***	=
Kocher													-	***	***	***	-	***	***	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	=	-	=	=	***
Modaresi																=	-	=	=	***
Pimas																	-	=	=	***
Poongunran																		-	-	-
Roman-Gomez																			=	***
Waser																				***
Zahid																				

Table A1. Significance of accuracy differences between system pairs for *gender* identification in the *English social media* corpus.

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		=	=	-	=	=	=	=	=	=	=	=	-	=	=	***	-	=	*	=
Ashraf			=	-	=	=	=	=	=	=	=	=	-	=	=	***	-	=	***	=
Bakkar				-	=	=	=	=	=	=	=	=	-	=	=	***	-	=	***	=
Bayot					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bilan						=	=	=	=	=	=	=	-	=	=	***	-	=	***	=
Bougiatiotis							=	=	=	=	=	=	-	=	=	***	-	=	**	=
Busger								=	=	=	=	=	-	=	=	***	-	=	**	=
Devalkeneer									=	=	=	=	-	=	=	***	-	=	***	=
Dichiu										=	=	=	-	=	=	***	-	=	**	=
Garciarena											=	=	-	=	=	***	-	=	***	=
Gencheva												=	-	=	=	***	-	=	**	=
Kocher													-	=	=	***	-	=	**	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	***	-	*	*	=
Modaresi																***	-	=	**	=
Pimas																	-	***	***	***
Poongunran																		-	-	-
Roman-Gomez																			***	=
Waser																				**
Zahid																				

Table A2. Significance of accuracy differences between system pairs for *age* identification in the *English social media* corpus.

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		***	=	-	=	=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Ashraf			***	-	***	***	=	=	***	***	=	=	-	***	***	***	-	***	***	=
Bakkar				-	=	=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Bayot					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bilan						=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Bougiatiotis							***	***	=	=	***	***	-	=	=	=	-	=	=	***
Busger								=	***	***	=	=	-	***	***	***	-	***	***	=
Devalkeneer									***	***	=	=	-	***	***	***	-	***	***	=
Dichiu										=	***	***	-	=	=	=	-	=	=	***
Garciarena											***	***	-	=	=	=	-	=	=	***
Gencheva												=	-	***	***	***	-	***	***	=
Kocher													-	***	***	***	-	***	***	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	=	-	=	=	***
Modaresi																=	-	=	=	***
Pimas																	-	=	=	***
Poongunran																		-	-	-
Roman-Gomez																			=	***
Waser																				***
Zahid																				

Table A3. Significance of accuracy differences between system pairs for *joint* identification in the *English social media* corpus.

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		***	=	-	=	=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Ashraf			***	-	***	***	=	=	***	***	=	=	-	***	***	***	-	***	***	=
Bakkar				-	=	=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Bayot					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bilan						=	***	***	=	=	***	***	-	=	=	=	-	=	=	***
Bougiatiotis							***	***	=	=	***	***	-	=	=	=	-	=	=	***
Busger								=	***	***	=	=	-	***	***	***	-	***	***	=
Devalkeneer									***	***	=	=	-	***	***	***	-	***	***	=
Dichiu										=	***	***	-	=	=	=	-	=	=	***
Garciarena											***	***	-	=	=	=	-	=	=	***
Gencheva												=	-	***	***	***	-	***	***	=
Kocher													-	***	***	***	-	***	***	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	=	-	=	=	***
Modaresi																=	-	=	=	***
Pimas																	-	=	=	***
Poongunran																		-	-	-
Roman-Gomez																			=	***
Waser																				***
Zahid																				

Table A4. Significance of accuracy differences between system pairs for *altogether* identification in the *English social media* corpus.

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		-	-	-	=	=	***	***	=	=	***	***	-	=	=	-	-	=	=	***
Ashraf			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bilan						=	***	***	=	=	***	***	-	=	=	-	-	=	=	***
Bougiatiotis							***	***	=	=	***	***	-	=	=	-	-	=	=	***
Busger								=	***	***	=	=	-	***	***	-	-	***	***	=
Devalkeneer									***	***	=	=	-	***	***	-	-	***	***	=
Dichiu										=	***	***	-	=	=	-	-	=	=	***
Garciarena											***	***	-	=	=	-	-	=	=	***
Gencheva												=	-	***	***	-	-	***	***	=
Kocher													-	***	***	-	-	***	***	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	-	-	=	=	***
Modaresi																-	-	=	=	***
Pimas																	-	-	-	-
Poongunran																		-	-	-
Roman-Gomez																			=	***
Waser																				***
Zahid																				

Table A5. Significance of accuracy differences between system pairs for *gender* identification in the *Spanish social media* corpus.

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		-	-	-	=	=	=	=	=	=	=	=	-	=	=	-	-	=	=	=
Ashraf			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bilan						=	=	=	=	=	=	=	-	=	=	-	-	=	=	=
Bougiatiotis							=	=	=	=	=	=	-	=	=	-	-	*	=	=
Busger								=	=	=	=	=	-	=	=	-	-	=	=	=
Devalkeneer									=	=	=	=	-	=	=	-	-	*	=	=
Dichiu										=	=	=	-	=	=	-	-	=	=	=
Garciarena											=	=	-	=	=	-	-	=	=	=
Gencheva												=	-	=	=	-	-	=	=	=
Kocher													-	=	=	-	-	=	=	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	-	-	=	=	=
Modaresi																-	-	=	=	=
Pimas																	-	-	-	-
Poongunran																		-	-	-
Roman-Gomez																			=	=
Waser																				=
Zahid																				

Table A6. Significance of accuracy differences between system pairs for *age* identification in the *Spanish social media* corpus.

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		-	-	-	=	=	***	***	=	=	**	***	-	=	=	-	-	=	=	**
Ashraf			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bilan						=	***	***	=	=	**	***	-	=	=	-	-	=	=	**
Bougiatiotis							***	***	=	=	**	***	-	=	=	-	-	=	=	**
Busger								=	***	***	=	=	-	***	***	-	-	***	***	=
Devalkeneer									***	***	=	=	-	***	***	-	-	***	***	=
Dichiu										=	**	***	-	=	=	-	-	=	=	**
Garciarena											**	***	-	=	=	-	-	=	=	**
Gencheva												=	-	**	**	-	-	**	**	=
Kocher													-	***	***	-	-	***	***	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	-	-	=	=	**
Modaresi																-	-	=	=	**
Pimas																	-	-	-	-
Poongunran																		-	-	-
Roman-Gomez																			=	**
Waser																				**
Zahid																				

Table A7. Significance of accuracy differences between system pairs for *joint* identification in the *Spanish social media* corpus.

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		-	-	-	=	=	***	***	=	=	**	***	-	=	=	-	-	=	=	**
Ashraf			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bilan						=	***	***	=	=	**	***	-	=	=	-	-	=	=	**
Bougiatiotis							***	***	=	=	**	***	-	=	=	-	-	=	=	**
Busger								=	***	***	=	=	-	***	***	-	-	***	***	=
Devalkeneer									***	***	=	=	-	***	***	-	-	***	***	=
Dichiu										=	**	***	-	=	=	-	-	=	=	**
Garciarena											**	***	-	=	=	-	-	=	=	**
Gencheva												=	-	**	**	-	-	**	**	=
Kocher													-	***	***	-	-	***	***	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	-	-	=	=	**
Modaresi																-	-	=	=	**
Pimas																	-	-	-	-
Poongunran																		-	-	-
Roman-Gomez																			=	**
Waser																				**
Zahid																				

Table A8. Significance of accuracy differences between system pairs for *altogether* identification in the *Spanish social media* corpus.

	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Agrawal		-	-	=	=	=	***	***	=	=	***	***	-	=	=	-	=	=	=	***
Ashraf			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot					=	=	***	***	=	=	***	***	-	=	=	-	=	=	=	***
Bilan						=	***	***	=	=	***	***	-	=	=	-	=	=	=	***
Bougiatiotis							***	***	=	=	***	***	-	=	=	-	=	=	=	***
Busger								=	***	***	=	=	-	***	***	-	***	***	***	=
Devalkeneer									***	***	=	=	-	***	***	-	***	***	***	=
Dichiu										=	***	***	-	=	=	-	=	=	=	***
Garciarena											***	***	-	=	=	-	=	=	=	***
Gencheva												=	-	***	***	-	***	***	***	=
Kocher													-	***	***	-	***	***	***	=
Markov														-	-	-	-	-	-	-
Modaresi(a)															=	-	=	=	=	***
Modaresi																-	=	=	=	***
Pimas																	-	-	-	-
Poongunran																		=	=	***
Roman-Gomez																			=	***
Waser																				***
Zahid																				

Table A9. Significance of accuracy differences between system pairs for *gender* identification in the *Dutch social media* corpus.

Appendix B Pairwise Comparison of all Systems in blogs

For all subsequent tables, the significance levels are encoded as follows:

Symbol	Significance Level		
-		∼	not evaluated
=	$p > 0.05$	∼	not significant
*	$0.05 \geq p > 0.01$	∼	significant
**	$0.01 \geq p > 0.001$	∼	very significant
***	$p \leq 0.001$	∼	highly significant

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		***	=	***	***	***	***	=	=	=	***	***	=	=	***	***	***	***	***	***	***	=
Agrawal			***	=	=	=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Ashraf				***	***	***	***	=	=	=	***	***	=	=	***	***	***	***	***	***	***	=
Bakkar					=	=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bayot						=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bilan							=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bougiatiotis								***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Busger									=	=	***	***	=	=	***	***	***	***	***	***	***	=
Deneva										=	***	***	=	=	***	***	***	***	***	***	***	=
Devalkeneer											***	***	=	=	***	***	***	***	***	***	***	=
Dichiu												=	***	***	=	=	=	=	=	=	=	***
Garciarena													***	***	=	=	=	=	=	=	=	***
Gencheva														=	***	***	***	***	***	***	***	=
Kocher															***	***	***	***	***	***	***	=
Markov																=	=	=	=	=	=	***
Modaresi(a)																	=	=	=	=	=	***
Modaresi																		=	=	=	=	***
Pimas																			=	=	=	***
Poongunran																				=	=	***
Roman-Gomez																					=	***
Waser																						***
Zahid																						

Table B1. Significance of accuracy differences between system pairs for *gender* identification in the *English blogs* corpus.

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		=	=	=	=	=	*	**	=	*	=	=	=	=	=	=	*	=	=	=	=	=
Agrawal			=	=	=	=	**	**	=	=	=	=	=	=	=	=	*	=	=	=	=	=
Ashraf				=	=	=	**	**	=	*	=	=	=	=	=	=	**	=	=	=	=	=
Bakkar					=	=	**	**	=	=	=	=	=	=	=	=	**	=	=	=	=	=
Bayot						=	*	**	=	=	=	=	=	=	=	=	*	=	=	=	=	=
Bilan							*	*	=	=	=	=	=	=	=	=	=	=	=	=	=	=
Bougiatiotis								=	*	=	=	=	*	*	=	=	=	**	**	**	=	=
Busger									**	=	**	**	**	*	*	=	=	**	***	**	=	**
Deneva										=	=	=	=	=	=	=	*	=	=	=	=	=
Devalkeneer											=	=	=	=	=	=	=	*	**	=	=	=
Dichiu												=	=	=	=	=	=	=	*	=	=	=
Garciarena													=	=	=	=	=	=	=	=	=	=
Gencheva														=	=	=	*	=	=	=	=	=
Kocher															=	=	=	=	=	=	=	=
Markov																=	=	=	*	=	=	=
Modaresi(a)																	=	=	**	=	=	=
Modaresi																		*	**	**	=	=
Pimas																			=	=	=	=
Poongunran																				=	*	=
Roman-Gomez																					=	=
Waser																						=
Zahid																						

Table B2. Significance of accuracy differences between system pairs for *age* identification in the *English blogs* corpus.

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		***	=	***	***	***	***	**	=	=	***	***	=	=	***	***	***	***	***	***	***	=
Agrawal			***	=	=	=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Ashraf				***	***	***	***	*	=	=	***	***	=	=	***	***	***	***	***	***	***	=
Bakkar					=	=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bayot						=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bilan							=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bougiatiotis								***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Busger									*	=	***	***	=	=	***	***	***	***	***	***	***	*
Deneva										=	***	***	=	=	***	***	***	***	***	***	***	=
Devalkeneer											***	***	=	=	***	***	***	***	***	***	***	=
Dichiu												=	***	***	=	=	=	=	=	=	=	***
Garciarena													***	***	=	=	=	=	=	=	=	***
Gencheva														=	***	***	***	***	***	***	***	=
Kocher															***	***	***	***	***	***	***	=
Markov																=	=	=	=	=	=	***
Modaresi(a)																	=	=	=	=	=	***
Modaresi																		=	=	=	=	***
Pimas																			=	=	=	***
Poongunran																				=	=	***
Roman-Gomez																					=	***
Waser																						***
Zahid																						

Table B3. Significance of accuracy differences between system pairs for *joint* identification in the *English blogs* corpus.

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		***	=	***	***	***	***	**	=	=	***	***	=	=	***	***	***	***	***	***	***	=
Agrawal			***	=	=	=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Ashraf				***	***	***	***	*	=	=	***	***	=	=	***	***	***	***	***	***	***	=
Bakkar					=	=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bayot						=	=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bilan							=	***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Bougiatiotis								***	***	***	=	=	***	***	=	=	=	=	=	=	=	***
Busger									*	=	***	***	=	=	***	***	***	***	***	***	***	*
Deneva										=	***	***	=	=	***	***	***	***	***	***	***	=
Devalkeneer											***	***	=	=	***	***	***	***	***	***	***	=
Dichiu												=	***	***	=	=	=	=	=	=	=	***
Garciarena													***	***	=	=	=	=	=	=	=	***
Gencheva														=	***	***	***	***	***	***	***	=
Kocher															***	***	***	***	***	***	***	=
Markov																=	=	=	=	=	=	***
Modaresi(a)																	=	=	=	=	=	***
Modaresi																		=	=	=	=	***
Pimas																			=	=	=	***
Poongunran																				=	=	***
Roman-Gomez																					=	***
Waser																						***
Zahid																						

Table B4. Significance of accuracy differences between system pairs for *altogether* identification in the *English blogs* corpus.

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		***	-	-	***	***	***	**	**	=	***	***	=	=	***	***	***	-	***	***	***	=
Agrawal			-	-	=	=	=	***	***	***	=	=	***	***	=	=	=	-	=	=	=	***
Ashraf				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot						=	=	***	***	***	=	=	***	***	=	=	=	-	=	=	=	***
Bilan							=	***	***	***	=	=	***	***	=	=	=	-	=	=	=	***
Bougiatiotis								***	***	***	=	=	***	***	=	=	=	-	=	=	=	***
Busger									=	=	***	***	=	=	***	***	***	-	***	***	***	*
Deneva										*	***	***	*	=	***	***	***	-	***	***	***	**
Devalkeneer											***	***	=	=	***	***	***	-	***	***	***	=
Dichiu												=	***	***	=	=	=	-	=	=	=	***
Garciarena													***	***	=	=	=	-	=	=	=	***
Gencheva														=	***	***	***	-	***	***	***	=
Kocher															***	***	***	-	***	***	***	=
Markov																=	=	-	=	=	=	***
Modaresi(a)																	=	-	=	=	=	***
Modaresi																		-	=	=	=	***
Pimas																			-	-	-	-
Poongunran																				=	=	***
Roman-Gomez																					=	***
Waser																						***
Zahid																						

Table B5. Significance of accuracy differences between system pairs for *gender* identification in the *Spanish blogs* corpus.

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		*	-	-	*	*	=	**	=	*	*	*	=	=	*	**	**	-	=	=	=	*
Agrawal			-	-	=	=	=	=	*	=	=	=	*	=	=	=	=	-	***	*	=	=
Ashraf				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot						=	=	=	*	=	=	=	*	=	=	=	=	-	***	*	=	=
Bilan							=	=	=	=	=	=	=	=	=	=	=	-	***	*	=	=
Bougiatiotis								*	=	=	=	=	=	=	=	=	*	-	=	=	=	=
Busger									*	=	=	=	*	=	=	=	=	-	***	**	=	=
Deneva										*	*	=	=	=	=	=	**	-	*	=	=	*
Devalkeneer											=	=	*	=	=	=	=	-	***	*	=	=
Dichiu												=	*	=	=	=	=	-	***	*	=	=
Garciarena													=	=	=	=	=	-	**	=	=	=
Gencheva														=	=	=	***	-	=	=	=	*
Kocher															=	=	=	-	=	=	=	=
Markov																=	=	-	**	=	=	=
Modaresi(a)																	=	-	**	*	=	=
Modaresi																		-	***	*	=	=
Pimas																			-	-	-	-
Poongunran																				=	**	**
Roman-Gomez																					=	=
Waser																						=
Zahid																						

Table B6. Significance of accuracy differences between system pairs for *age* identification in the *Spanish blogs* corpus.

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		=	-	-	=	=	=	***	*	*	=	=	=	=	=	=	=	-	=	=	=	=
Agrawal			-	-	=	=	=	***	***	***	=	=	***	***	=	=	=	-	=	=	=	***
Ashraf				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot						=	=	***	***	***	=	=	***	**	=	=	=	-	=	=	=	***
Bilan							=	***	***	***	=	=	***	**	=	=	=	-	=	=	=	***
Bougiatiotis								***	***	***	=	=	***	***	=	=	=	-	=	=	=	***
Busger									=	=	***	***	*	*	***	***	***	-	***	***	***	*
Deneva										=	***	***	=	=	***	***	***	-	***	***	***	=
Devalkeneer											***	***	=	=	***	***	***	-	***	***	***	=
Dichiu												=	***	**	=	=	=	-	=	=	=	***
Garciarena													***	***	=	=	=	-	=	=	=	***
Gencheva														=	***	***	***	-	***	***	***	=
Kocher															***	**	***	-	**	***	***	=
Markov																=	=	-	=	=	=	***
Modaresi(a)																	=	-	=	=	=	***
Modaresi																		-	=	=	=	***
Pimas																			-	-	-	-
Poongunran																				=	=	***
Roman-Gomez																					=	***
Waser																						***
Zahid																						

Table B7. Significance of accuracy differences between system pairs for *joint* identification in the *Spanish blogs* corpus.

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		=	-	-	=	=	=	***	*	*	=	=	=	=	=	=	=	-	=	=	=	=
Agrawal			-	-	=	=	=	***	***	***	=	=	***	**	=	=	=	-	=	=	=	***
Ashraf				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot						=	=	***	***	***	=	=	***	**	=	=	=	-	=	=	=	***
Bilan							=	***	***	***	=	=	***	**	=	=	=	-	=	=	=	***
Bougiatiotis								***	***	***	=	=	***	***	=	=	=	-	=	=	=	***
Busger									=	=	***	***	*	*	***	***	***	-	***	***	***	*
Deneva										=	***	***	=	=	***	***	***	-	***	***	***	=
Devalkeneer											***	***	=	=	***	***	***	-	***	***	***	=
Dichiu												=	***	**	=	=	=	-	=	=	=	***
Garciarena													***	***	=	=	=	-	=	=	=	***
Gencheva														=	***	***	***	-	***	***	***	=
Kocher															**	***	***	-	**	**	**	=
Markov																=	=	-	=	=	=	***
Modaresi(a)																	=	-	=	=	=	***
Modaresi																		-	=	=	=	***
Pimas																			-	-	-	-
Poongunran																				=	=	***
Roman-Gomez																					=	***
Waser																						***
Zahid																						

Table B8. Significance of accuracy differences between system pairs for *altogether* identification in the *Spanish blogs* corpus.

	Aceituno	Agrawal	Ashraf	Bakkar	Bayot	Bilan	Bougiatiotis	Busger	Deneva	Devalkeneer	Dichiu	Garciarena	Gencheva	Kocher	Markov	Modaresi(a)	Modaresi	Pimas	Poongunran	Roman-Gomez	Waser	Zahid
Aceituno		***	-	-	***	***	***	=	***	=	***	***	=	=	***	***	***	-	***	***	***	-
Agrawal			-	-	=	=	***	***	***	***	=	=	***	***	=	=	=	-	=	=	=	-
Ashraf				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bakkar					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bayot						=	***	***	***	***	=	=	***	***	=	=	=	-	=	=	=	-
Bilan							***	***	***	***	=	=	***	***	=	=	=	-	=	=	=	-
Bougiatiotis								***	***	***	***	***	***	***	***	***	***	-	***	***	***	-
Busger									***	=	***	***	=	=	***	***	***	-	***	***	***	-
Deneva										**	***	***	***	***	***	***	***	-	***	***	***	-
Devalkeneer											***	***	=	=	***	***	***	-	***	***	***	-
Dichiu												=	***	***	=	=	=	-	=	=	=	-
Garciarena													***	***	=	=	=	-	=	=	=	-
Gencheva														=	***	***	***	-	***	***	***	-
Kocher															***	***	***	-	***	***	***	-
Markov																=	=	-	=	=	=	-
Modaresi(a)																	=	-	=	=	=	-
Modaresi																		-	=	=	=	-
Pimas																			-	-	-	-
Poongunran																				=	=	-
Roman-Gomez																					=	-
Waser																						-
Zahid																						

Table B9. Significance of accuracy differences between system pairs for *gender* identification in the *Dutch blogs* corpus.