Overview of the 2nd Author Profiling Task at PAN 2014

Francisco Rangel[1,2] Paolo Rosso[2] Irina Chugur[3]
Martin Potthast[4] Martin Trenkmann[4] Benno Stein[4]
Ben Verhoeven[5] Walter Daelemans[5]

[1]Autoritas Consulting, S.A., Spain
[2]Natural Language Engineering Lab, Universitat Politècnica de València, Spain
[3]Universidad Nacional de Educación a Distancia, Madrid, Spain
[4]Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany
[5]CLiPS - Computational Linguistics Group, University of Antwerp, Belgium

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

Abstract This overview presents the framework and the results for the Author Profiling task at PAN 2014. Objective of this year is the analysis of the adaptability of the detection approaches when given different genres. For this purpose a corpus with four different parts (subcorpora) has been compiled: social media, Twitter, blogs, and hotel reviews. The construction of the Twitter subcorpus happened in cooperation with RepLab in order to investigate also a reputational perspective. Altogether, the approaches of 10 participants are evaluated.

1 Introduction

Though the enormous impact of social media on our daily life, we observe a lack of information about those who create the contents. In this regard, author profiling tries to determine the gender, age, native language, or personality type of authors by analysing their published texts. Author profiling is of growing importance: E.g., from a marketing viewpoint, companies may be interested in knowing the demographics of their target group in order to achieve a better market segmentation; from a forensic viewpoint, determining the linguistic profile of a person who wrote a "suspicious text"' may provide valuable background information.

In the Author Profiling task at PAN 2013,[1] the identification of age and gender relied on a large corpus collected from social media [28]. This year, in PAN 2014,[2] we continue focusing on age and gender aspects but, in addition, compiled a corpus of four different genres, namely social media, blogs, Twitter, and hotel reviews. Except for the hotel review subcorpus, which is available in English only, all documents are provided in both English and Spanish. Note that most of the existing research in computational linguistics [3] and social psychology [26] focuses on the English language, and the question is whether the observed relations pertain to other languages as well.

The remainder of this paper is organised as follows. Section 2 covers the state of the art, Section 3 describes the corpus and evaluation measures, and Section 4 presents

[1] http://webis.de/research/events/pan-13/pan13-web/author-profiling.html
[2] http://webis.de/research/events/pan-14/pan14-web/author-profiling.html

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 *et al.* [27] connected language use with personality traits, studying how the variation of linguistic characteristics in a text can provide information regarding the gender and age of its author. Argamon *et al.* [3] analysed formal written texts extracted from the British National Corpus, combining function words with part-of-speech features and achieving approximately 80% accuracy in gender prediction. Other researchers (Holmes and Meyerhoff [14], Burger and Henderson [5]) have also investigated how to obtain age and gender information from formal texts.

With the rise of the social media, the focus is on other kind of writings, more colloquial, less structured and formal, like blogs or fora. Koppel *et al.* [15] studied the problem of automatically determining an author's gender by proposing combinations of simple lexical and syntactic features, and achieving approximately 80% accuracy. Schler *et al.* [29] studied the effect of age and gender in the writing style in blogs; they gathered over 71,000 blogs and obtained a set of stylistic features like non-dictionary words, parts-of-speech, function words and hyperlinks, combined with content features, such as word unigrams with the highest information gain. They obtained an accuracy of about 80% for gender identification and about 75% for age identification. They modeled age in three classes: 10s (13-17), 20s (23-27) and 30s (33-47). They demonstrated that language features in blogs correlates with age, as reflected in, for example, the use of prepositions and determiners. Goswami *et al.* [12] added some new features as slang words and the average length of sentences, improving accuracy to 80.3% in age group identification and to 89.2% in gender detection.

It is to be noted that the previously described studies were conducted with texts of at least 250 words. The effect of data size is known, however, to be an important factor in machine learning algorithms of this type. In fact, Zhang and Zhang [34] experimented with short segments of blog post, specifically 10,000 segments with 15 tokens per segment, and obtained 72.1% accuracy for gender prediction, as opposed to more than 80% in the previous studies. Similarly, Nguyen *et al.* [23] studied the use of language and age among Dutch Twitter users, where the documents are really short, with an average length of less than 10 terms. They modelled age as a continuous variable (as they had previously done in [22]), and used an approach based on logistic regression. They also measured the effect of the gender in the performance of age identification, considering both variables as inter-dependent, and achieved correlations up to 0.74 and mean absolute errors between 4.1 and 6.8 years.

One common problem when investigating author profiling is the need to obtain labelled data for the authors, to obtain their age and gender. Studies in classical literature deal with a small number of well-known authors, where manual labelling can easily be applied. However for the dimensions of the actual social media data this is a more

difficult task, which should be automated. In some cases, researchers manually label the collection [23] with some risk of bias. In other cases, as in the vast majority of the aforementioned studies, researchers took into account information provided by the authors themselves. For example, in blog platforms, the contributors self-specify their profiles. This is the case for Peersman *et al.* [25] who retrieved a dataset from Netlog,[3] where authors report their gender and exact age, and Koppel *et al.* [15], who retrieved the dataset from Blogspot.[4] This is likely to introduce some noise to the evaluation set, but it also reflects the realistic state of the available data.

The task of obtaining author profiles has an emerging interest in the scientific community, as can be seen in the number of related tasks around the topic arisen the two last years: *a*) the shared task on Native Language Identification at BEA-8 Workshop at NAACL-HT 2013;[5] *b*) the task on Computational Personality Recognition (WCPR) at ICWSM 2013[6] and at ACM Multimedia 2014,[7] and; *c*) the task on Author Profiling at PAN 2013 and PAN 2014.

With respect to the task on Author Profiling at PAN 2013 [28], most of the participants used combinations of style-based features such as frequency of punctuation marks, capital letters, quotations, and so on, together with POS tags and content-based features such as Latent Semantic Analysis, bag-of-words, TF-IDF, dictionary-based words, topic-based words, and so on. It is worth mentioning the usage of second order representations based on relationships between documents and profiles by the winner of the PAN-AP 2013 task [16] and the use of collocations for the winner of the English task [21].

Last but not least, the interest in different author profile aspects is evident also in the Kaggle platform,[8] where companies and research departments shared their needs and independent researchers joined challenges as Psychopathy Prediction Based on Twitter Usage;[9] Personality Prediction Based on Twitter Stream;[10] or Gender Prediction from Handwriting.[11] This shows the rise of interest from the industry in author profiling.

3 Evaluation Framework

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

3.1 Corpus

In order to study how the different author profiling approaches apply to different genres, we have built a corpus with four different genres: social media, blogs, Twitter, and hotel

[3] http://www.netlog.com
[4] http://blogspot.com
[5] https://sites.google.com/site/nlisharedtask2013/
[6] http://mypersonality.org/wiki/doku.php?id=wcpr13
[7] https://sites.google.com/site/wcprst/home/wcpr14
[8] http://www.kaggle.com/
[9] http://www.kaggle.com/c/twitter-psychopathy-prediction
[10] http://www.kaggle.com/c/twitter-personality-prediction
[11] http://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting

reviews. The respective subcorpora cover English and Spanish, with the exception of the hotel reviews, which have been provided in English only. The corpus documents are encoded as XML files, one per author, with the contents between <document> tags. The author is labeled with age and gender information. For labeling age, instead of the three age classes *a*) 10s (13-17); *b*) 20s (23-27); *c*) 30s (33-47) used in PAN-AP 2013, this year we opted for modelling age in a more fine-grained way and considered the following classes: *a*) 18-24; *b*) 25-34; *c*) 35-49; *d*) 50-64; *e*) 65+ .

As in the previous edition, each subcorpus was split into three parts for training, early birds, and test respectively.

Social Media We have built the social media subcorpus by selecting a part of the PAN-AP-13 corpus. We have selected those authors with an average number of words in their posts greater than 100. We also manually reviewed the documents in order to remove those authors who seem to be fake profiles such as bots, for example, authors selling the same product (e.g., mobiles, ads) in most of their posts or authors with a high number of text reuse (e.g., teenagers sharing poetry or homework). The final distribution of the number of authors is shown in Table 1. The social media subcorpus is balanced by gender, so the number of authors per gender is one-half.

Table 1. Distribution of social media with respect to age classes per language.

	Training		Early birds		Test	
	English	Spanish	English	Spanish	English	Spanish
18-24	1550	330	140	30	680	150
25-34	2098	426	180	36	900	180
35-49	2246	324	200	28	980	138
50-64	1838	160	160	14	790	70
65+	14	32	12	14	26	28
Σ	7746	1272	692	122	3376	566

Blogs The objective of collecting blogs is to build a gold standard for author profiling in this specific genre. To achieve this objective, we manually selected and annotated the documents. Firstly, we looked for public LinkedIn profiles which share a personal blog URL. We verified that the blog exists, it is written in one of the languages we are interested in (English or Spanish) and it is updated only by one person and this person is easily identifiable. We discarded organizational blogs when we were not sure that the blog was updated by the person identified in the LinkedIn profile. Secondly, we looked for age information. In some cases the birth date is published in the user's profile. But in most cases it is not so we looked for degree starting date in the education section. We used the information shown in Table 2 to figure out the age range. We discarded users whose education dates were not clear. Thirdly, if we could figure out the age, we identified the gender by the user's photography and name. Again, for those cases where the gender information was not clear, we discarded the user. Finally, this process was done by two independent annotators and a third one decided in case of disagreement. For each blog, we provided up to 25 posts. We provided contents obtained from the RSS feed but we allow users to download the full text from the permalink.

Table 2. Age range by 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

The final distribution of the number of authors is shown in Table 3. The blogs subcorpus is balanced by gender, so the number of authors per gender is half.

Table 3. Distribution of blogs with respect to age classes per language.

	Training		Early birds		Test	
	English	Spanish	English	Spanish	English	Spanish
18-24	6	4	4	2	10	4
25-34	60	26	6	4	24	12
35-49	54	42	8	4	32	26
50-64	23	12	4	2	10	10
65+	4	4	2	2	2	2
Σ	147	88	24	14	78	56

Twitter We manually selected and annotated the documents, following the same methodology as for the blogs. We built this subcorpus in collaboration with RepLab[12] where the main goal of author profiling—viewed in the context of reputation monitoring on Twitter—is to decide how influential a given user is in the domain which the entity under study belongs to. This includes determining the type of author (e.g., journalist, stakeholder, professional) and his degree of influence on opinions within the domain. For the shared PAN-RepLab author profiling task, 131 Twitter profiles from several domains (energy, environmental, banking, automotive, and Corporate Social Responsibility sectors) were annotated with age and gender. The profiles were selected from the RepLab 2013 corpus and from a list of influential authors provided by the online division of a leading Public Relations consultancy (Llorente & Cuenca).[13] Note that balancing the list of profiles by age and gender turned out to be a challenging task, because influential Twitter authors in the considered economic domains tend to be male and of quite a narrow age range (35-49). In addition to age and gender, tweets in RepLab were manually tagged by reputation experts with *a*) type of author and; *b*) opinion-maker labels (Influencer, Non-influencer, and Undecidable).

For more details on the RepLab 2014 author profiling data set please refer to [2]. Due to Twitter terms of service, we provided the tweets URLs so that participants could download them. For each Twitter profile, we provided up to 1000 tweets. The final

[12] http://nlp.uned.es/replab2014

[13] http://www.llorenteycuenca.com/

distribution of the number of authors is shown in Table 4. The Twitter subcorpus is balanced by gender, so half of the authors are male and the other half are female.

Table 4. Distribution of Twitter with respect to age classes per language.

	Training		Early birds		Test	
	English	Spanish	English	Spanish	English	Spanish
18-24	20	12	2	2	12	4
25-34	88	42	6	4	56	26
35-49	130	86	16	12	58	46
50-64	60	32	4	6	26	12
65+	8	6	2	2	2	2
Σ	306	178	30	26	154	90

Hotels Reviews To study the applicability of author profiling approaches to the review genre, we have compiled the Webis-TripAd-13 corpus, a large subset of hotel reviews from the PAN 2014 author profiling evaluation corpus. The corpus has been carefully constructed to ensure its quality with regard to text cleanliness and annotation accuracy.

The Webis-TripAd-13 corpus is derived from another corpus that was originally used for aspect-level rating prediction [31].[14] The original corpus was crawled from the hotel review site TripAdvisor[15] in the period of one month from mid February to mid March 2009, and contains 235 793 reviews about 1,850 different hotels. Each review comprises its author's user name, the review text, and the date the review was written. In addition, there are seven numerical aspect ratings and an overall rating score assigned by the user, which serve as ground-truth for aspect-level rating prediction or sentiment analysis tasks in general. However, the original dataset does not feature age and gender annotations.

In order to make this dataset applicable to author profiling and to ensure its quality, we applied the following four post-processing steps: first, we removed short reviews of less than 10 words which were found to be malformed reviews due to parsing errors. Second, we removed reviews whose text was not found to be English according to a language detector. Third, since the original dataset does not provide any age and gender information, we compiled a list of user names who submitted the reviews and crawled the corresponding user profiles from the TripAdvisor website. Fourth, given this metadata, we discarded all reviews written by authors whose age and gender was not given on their user profile or whose user profile was inactive. Moreover, to ensure data quality, we reviewed user profiles and reviews with regard to sanity (i.e., whether the information given made sense). The final Webis-TripAd-13 corpus contains 58 101 reviews and covers six age classes. The distribution of reviews across these classes is shown in columns 3 and 4 of Table 5.[16]

[14] http://times.cs.uiuc.edu/~wang296/Data

[15] http://www.tripadvisor.com

[16] This version of the corpus has been released at: http://www.webis.de/research/corpora

To match the requirements of PAN's author profiling evaluation corpus, we unified the Webis-TripAd-13 corpus accordingly: to obtain a nearly uniform age class distribution, we sampled 700 authors from each of the three major classes (25–34, 35–49, 50–64). For the two minor classes (18–24, 65+), however, the number of authors available was limited by the size of the smaller age class, so that 254 authors (18–24) and 547 authors (65+) remained, respectively. Class 13–17 was discarded completely since the number of available authors was found to be not representative for evaluation purposes. The final distribution of the subset of the Webis-TripAd-13 corpus that forms part of the PAN author profiling evaluation corpus is shown in Table 5, column 7–8.

Table 5. Distribution of reviews with respect to age and gender classes.

		Webis-TripAd-13		PAN 2014 training set		PAN 2014 test set	
Gender	Age	# authors	# reviews	# authors	# reviews	# authors	# reviews
female	13-17	23	23	-	-	-	-
	18-24	656	741	180	208	74	84
	25-34	7517	9504	500	651	200	247
	35-49	10554	13552	500	659	200	255
	50-64	5850	7449	500	617	200	242
	65+	547	682	400	494	147	188
male	13-17	22	25	-	-	-	-
	18-24	254	314	180	228	74	86
	25-34	3816	5144	500	700	200	250
	35-49	8586	12044	500	707	200	302
	50-64	5413	7229	500	669	200	268
	65+	1079	1394	400	520	147	178

3.2 Performance measures

For evaluating participants' approaches we have used accuracy. More specifically, we calculated the ratio between the number of authors correctly predicted by the total number of authors. We calculated separately accuracy for each subcorpus, language, gender, and age class. Moreover, we combined accuracy for the joint identification of age and gender. The final score used to rank the participants is the average for the combined accuracies for each subcorpus and language.

We computed statistical significance of performance differences between systems using approximate randomisation testing [24].[17] As noted by Yeh [33], 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 and with $p < 0.05$, we consider the systems to be significantly

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

different from each other. The complete set of statistical significance tests is illustrated in Appendix A.

In case of age identification we also measured the average and standard deviation of the distance between the predicted and the truth class. We define the distance between classes as the number of hops between them, with the maximum distance equal to 4 in case of the most distant ones (18-24 and 65+). In case the participant did not provide a prediction, we added 1 to the maximum distance, penalising this missing value with a distance of 5. We also calculated the total time needed to process the test data, in order to investigate the applicability in a real world.

3.3 Software Submissions

We continue to invite software submissions instead of run submissions for the second time. Within 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, for example, on future evaluation corpora. To facilitate software submissions, we develop the TIRA experimentation platform [9, 10], which makes handling software submissions at scale as simple as handling run submissions. Using TIRA, participants deploy their software into virtual machines at our site, which allows us to keep them in a running state [11].

4 Overview of the Submitted Approaches

Ten teams participated in the Author Profiling task. Eight of them submitted the notebook paper, a further one (liau14) provided us with a description of the approach, and castillojuarez14 did not comment on any change with respect to their last year's system [1].

Pre-processing. Various participants cleaned the HTML and XML to obtain plain text [18, 19, 4, 13, 32]. One participant [13] removed URLs, user mentions and hashtags from the Twitter texts. In [4], participants carried out case conversion, deleted invalid characters and multiple white spaces, and similarly in [32] where the participants also escaped invalid characters. Only in [30] and [32] participants performed tokenisation, whereas in [32] they studied the effect of subset selection, and in [19] they tried to delete spam bots by deleting contents with high percentage of the % character.

Features. Many participants [20, 19, 13, 4, 32, 18] and (liau14) considered different kinds of stylistic features. For example frequencies of different punctuation signs were used in [13, 20, 4, 18, 32], size of sentences, words that appear once and twice or the use of deflections in [20], the number of characters, words and sentences in [32]. In [19] participants measured the number of posts per user, the frequency of capital letters and capital words, whereas in [32] participants measured the correctness, cleanliness and diversity of the texts. Only in [32] and [19] participants took advantage of the HTML information, using the occurrence of tags such as *img*, *href* or *br*. Different readability features where used in [20, 19, 13, 4, 32]. For example, Automated Readability

Index [19, 13], Coleman-Liau Index [19, 13], Rix Readability Index [19, 13], Gunning Fox Index [13], Flesch-Kinkaid [32]. A lexical analysis was carried out in [20] and [13], where participants employed parts-of-speech as features together with the identification of proper nouns or words with character flooding (e.g., hellooooo). The occurrence of emoticons was used in [18], [19] and liau14.

With respect to content features, in [30, 18] and (liau14) participants modeled the language with n-grams or bag-of-words. In [20] they extracted topic words such as *money*, *home*, *smartphone*, *games*, *sports*, *job*, *marketing*, etc. In [19] participants used MRC and LIWC features to extract frequency of words related to different psycholinguistic concepts such as *familiarity*, *concreteness*, *imagery*, *motion*, *emotion*, *religion*, and so on. Some participants used dictionaries to differentiate words per subcorpus and class [4], identify lexical errors [19], foreign words [13] or specific phrases such as *my husband* or *my wife* [19] and liau14.

Specific features were used in [32], where participants obtained features employed in information retrieval (IR) such as the cosine similarity or the Okapi BM25. Finally, in [19] participants estimated the sentiment of the sentences and in [17] participants used a second order representation based on relationships among terms, documents, profiles and subprofiles.

Classification approaches. All the participants approached the task as a machine learning task. For example, logistic regression was used in [18] and liau14, and also in [32] where participants used a different algorithm per subcorpus, for instance logic boost, rotation forest, multi-class classifier, multilayer perceptron and simple logistic. In [30] participants used multinominal Naïve Bayes, in [17] libLINEAR, in [13] random forests, in [19] support vector machines and in [20] decision tables. In [4] participants implemented their own frequency-based prediction function.

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 some feedback. There were 7 early bird submissions and eventually 10 for final evaluation. We show results separately for the evaluation in each corpus part and for each language. Results are given in accuracy of identification of age, gender, as well as the joint identification of age and gender. Results for early birds are shown in Tables 6 - 9, whereas final results are shown in Tables 10 to 13. In case of final evaluation, a baseline was provided for comparison purposes. This baseline considered the 1 000 most frequent character trigrams. Some participants did not run their systems on any of the subcorpora.

As can be seen in the early bird results, the best ones were obtained for Twitter, both in English and Spanish, with no big differences between the two languages. In case of blogs, there are similar results for gender identification, but for age and joint identification the best results were obtained on the Spanish partition. The English blogs subcorpus is the one with the lowest results in age and joint identification, together with social media in English and hotel reviews in joint identification. For social media, the results are better in Spanish than in English for all the predictions. Spanish social media got one of the highest accuracies in gender identification, together with hotel

reviews and Twitter texts. With respect to hotel reviews, gender accuracies are close to Twitter, but age and joint identification belong to the lowest among all subcorpora. The highest values were obtained by shrestha14 [18] on Spanish Twitter with 0.8846 in gender identification, 0.6923 in age identification and 0.6154 in joint identification of both age and gender.

Table 6. Evaluation results for early birds in social media in terms of accuracy on English (left) and Spanish (right) texts.

English				**Spanish**			
Team	Joint	Gender	Age	Team	Joint	Gender	Age
liau14	0.2153	0.5390	0.3728	shrestha14	0.3033	0.6803	0.4016
shrestha14	0.2009	0.5332	0.3627	liau14	0.2787	0.7295	0.4262
lopezmonroy14	0.1893	0.5332	0.3338	lopezmonroy14	0.2377	0.6639	0.3689
castillojuarez14	0.1517	0.5231	0.3035	marquardt14	0.1639	0.6803	0.2705
marquardt14	0.1517	0.5260	0.2717	baker14	0.1557	0.5000	0.3115
ashok14	0.1272	0.5072	0.2558	castilloJuarez14	0.0656	0.4754	0.2049
baker14	0.1257	0.5000	0.2529	ashok14	-	-	-

Table 7. Evaluation results for early birds in blogs in terms of accuracy on English (left) and Spanish (right) texts.

English				**Spanish**			
Team	Joint	Gender	Age	Team	Joint	Gender	Age
lopezmonroy14	0.2083	0.6250	0.2500	lopezmonroy14	0.3571	0.5000	0.4286
liau14	0.1667	0.5000	0.2083	marquardt14	0.2857	0.6429	0.3571
ashok14	0.1667	0.4583	0.1667	shrestha14	0.2857	0.5714	0.4286
shrestha14	0.1667	0.5417	0.2500	castillojuarez14	0.2143	0.5000	0.3571
marquardt14	0.1250	0.5417	0.2500	baker14	0.1429	0.5000	0.2857
castillojuarez14	0.0833	0.5833	0.2500	liau14	0.0714	0.4286	0.2857
baker14	0.0417	0.5000	0.2083	ashok14	-	-	-

Table 8. Evaluation results for early birds in Twitter in terms of accuracy on English (left) and Spanish (right) texts.

English				**Spanish**			
Team	Joint	Gender	Age	Team	Joint	Gender	Age
lopezmonroy14	0.5333	0.7667	0.6333	shrestha14	0.6154	0.8846	0.6923
shrestha14	0.4000	0.7333	0.4333	lopezmonroy14	0.5385	0.7692	0.5769
liau14	0.3667	0.6667	0.5667	liau14	0.3846	0.6923	0.5385
marquardt14	0.3000	0.5667	0.5333	marquardt14	0.3846	0.7692	0.5000
baker14	0.2667	0.5333	0.5000	baker14	0.1923	0.5000	0.4615
ashok14	0.2333	0.5000	0.4667	ashok14	-	-	-
castillojuarez14	-	-	-	castillojuarez14	-	-	-

Table 9. Evaluation results for early birds in hotel reviews in terms of accuracy on English texts.

	English		
Team	Joint	Gender	Age
liau14	0.2622	0.7317	0.3415
lopezmonroy14	0.2500	0.6524	0.3720
shrestha14	0.2012	0.6280	0.2805
marquardt14	0.1585	0.5976	0.2561
ashok14	0.1220	0.5854	0.2317
baker14	0.1037	0.5427	0.2439
castillojuarez14	0.0854	0.4756	0.1951

As for the early birds, the best results in the final evaluation were achieved for Twitter. In this case gender identification accuracies are higher in English whereas age and joint identification are higher in Spanish. In any case, all the results are much lower than the early birds ones, where the size of the set was approximately 10%. With respect to the blogs, the best results in gender identification were achieved in English and for age identification in Spanish. Although the joint identification obtained similar values, in English there are more participants with higher results. The lowest accuracy for gender identification was reoprted for the Spanish blogs, with values very close to the random chance. These results are even worse than the early birds ones. Most of the participants obtained better results for English than in the early birds, except marquardt14 [19] who obtained worse results. Results in social media and hotel reviews are very similar to the early birds ones, probably caused by the large number of authors. The results for blogs are very similar to social media in case of age identification. The lowest results in joint identification were achieved in English social media and in hotel reviews, where furthermore the lowest results in age identification were obtained. The lowest results in gender identification were achieved in English blogs, with values very close to the random chance. On the contrary, the highest results for gender identification were achieved in hotel reviews and in Twitter. The high ranking of the baseline approach in hotel reviews is noteworthy, with values for gender identification of 0.6626 and a joint identification just in mid-ranking.

The highest effectiveness values were achieved by liau14 in gender identification on English Twitter (accuracy of 0.7338) and by shrestha14 [18] in age identification on Spanish Twitter (accuracy of 0.6111) as well as in joint identification on Spanish Twitter (accuracy of 0.4333). It is difficult to draw a correlation between approaches and results, but looking at the three highest accuracies per subcorpus and task (gender, age and joint identification), it seems that on overall simple content features such as bag-of-words or word n-grams achieve the best results. Similarly, bag-of-words used by liau14, word n-grams used by shrestha14 [18] and term vector model used by villenaroman14 [30] achieved the best results for almost all genres. Also noteworthy is the contribution of IR features used by weren14 [32] in all the identifications in English blogs, joint identification in English social media, age identification in Spanish Twitter, Spanish social media and hotel reviews, gender identification in Spanish blogs and joint identification in English social media. The mix of content and style features of marquardt14 [19] gave good results in gender identification in Spanish Twitter and in

the three identifications in Spanish blogs. The second ranking in gender identification in Spanish social media was obtained with the char n-grams baseline, but low rankings in the other subcorpora demonstrate that the use of character n-grams does not seem to be a good approach for author profiling in general. The overall best performance was obtained by lopezmonroy14 [17] employing second order representation based on terms. Table 14 shows the joint identification accuracies per subcorpus and their average.

Table 10. Evaluation results in social media in terms of accuracy on English (left) and Spanish (right) texts.

English				**Spanish**			
Team	Joint	Gender	Age	Team	Joint	Gender	Age
shrestha14	0.2062	0.5382	0.3652	liau14	0.3357	0.6837	0.4894
liau14	0.1952	0.5385	0.3605	shrestha14	0.2845	0.6449	0.4276
weren14	0.1914	0.5361	0.3489	lopezmonroy14	0.2809	0.6431	0.4523
villenaroman14	0.1905	0.5421	0.3581	weren14	0.2792	0.6307	0.4382
lopezmonroy14	0.1902	0.5237	0.3552	marquardt14	0.2102	0.6431	0.3445
castillojuarez14	0.1445	0.5053	0.2855	villenaroman14	0.1961	0.5724	0.3622
marquardt14	0.1428	0.5216	0.2701	baseline	0.1820	0.6555	0.2862
ashok14	0.1318	0.5198	0.2515	baker14	0.1678	0.5000	0.3445
baker14	0.1277	0.5012	0.2494	castillojuarez14	0.1254	0.4982	0.2509
mechti14	0.1244	0.5198	0.2355	mechti14	0.1060	0.5919	0.2191
baseline	0.0930	0.5074	0.1925	ashok14	-	-	-

Table 11. Evaluation results in blogs in terms of accuracy on English (left) and Spanish (right) texts.

English				**Spanish**			
Team	Joint	Gender	Age	Team	Joint	Gender	Age
lopezmonroy14	0.3077	0.6795	0.3974	lopezmonroy14	0.3214	0.5893	0.4821
villenaroman14	0.3077	0.6410	0.3974	marquardt14	0.2679	0.5179	0.4821
weren14	0.2949	0.6410	0.4615	shrestha14	0.2500	0.4286	0.4643
liau14	0.2692	0.6538	0.3462	baker14	0.2321	0.5000	0.4464
shrestha14	0.2308	0.5769	0.3846	liau14	0.2321	0.5000	0.4464
castillojuarez14	0.1795	0.5128	0.3333	villenaroman14	0.2321	0.5179	0.4643
ashok14	0.1282	0.4231	0.2564	mechti14	0.1786	0.5000	0.2857
baker14	0.1282	0.5000	0.2949	weren14	0.1786	0.5357	0.2500
marquardt14	0.1282	0.4615	0.2692	castillojuarez14	0.0893	0.4464	0.2679
baseline	0.0897	0.5769	0.1410	baseline	0.0536	0.5357	0.1607
mechti14	0.0897	0.5897	0.1795	ashok14	-	-	-

Table 12. Evaluation results in Twitter in terms of accuracy on English (left) and Spanish (right) texts.

English				Spanish			
Team	Joint	Gender	Age	Team	Joint	Gender	Age
lopezmonroy14	0.3571	0.7208	0.4935	shrestha14	0.4333	0.6556	0.6111
liau14	0.3506	0.7338	0.5065	lopezmonroy14	0.3444	0.6000	0.5333
shrestha14	0.3052	0.6688	0.4416	liau14	0.3222	0.6333	0.5000
villenaroman14	0.2078	0.5130	0.4156	marquardt14	0.3111	0.6111	0.5222
weren14	0.2013	0.5714	0.3312	weren14	0.2778	0.5333	0.5222
ashok14	0.1948	0.5000	0.3896	villenaroman14	0.2667	0.5444	0.5000
marquardt14	0.1948	0.5260	0.3766	baseline	0.2333	0.4778	0.4667
baker14	0.1688	0.5065	0.3377	baker14	0.2111	0.5000	0.4889
baseline	0.1494	0.5974	0.2792	mechti14	0.1444	0.5111	0.2222
mechti14	0.0584	0.5390	0.1104	ashok14	-	-	-
castilloJuarez14	-	-	-	castillojuarez14	-	-	-

Table 13. Evaluation results in hotel reviews in terms of accuracy on English texts.

English			
Team	Joint	Gender	Age
liau14	0.2564	0.7259	0.3502
lopezmonroy14	0.2247	0.6809	0.3337
shrestha14	0.2223	0.6687	0.3331
weren14	0.2211	0.6778	0.3343
villenaroman14	0.2199	0.6845	0.3143
baseline	0.1821	0.6626	0.2753
marquardt14	0.1437	0.5700	0.2436
baker14	0.1382	0.5292	0.2594
ashok14	0.1291	0.5189	0.2454
castillojuarez14	0.1236	0.5091	0.2418
mechti14	0.0451	0.5012	0.0901

In Table 14 joint identification accuracies per subcorpus and the average are shown. From this table we can infer that: *a*) the best results were obtained on Twitter maybe due to the higher number of documents (tweets) per author in comparison to the other genre and quite likely also to the spontaneous way people express themselves; *b*) the lowest results were achieved in English social media and hotel reviews, due to the lowest results in gender identification in the first case and age identification in the second one.

Table 14. Average results in terms of accuracy.

Ranking	Team	Average	Social Media		Blogs		Twitter		Reviews
			EN	ES	EN	ES	EN	ES	EN
1	lopezmonroy14	0.2895	0.1902	0.2809	0.3077	0.3214	0.3571	0.3444	0.2247
2	liau14	0.2802	0.1952	0.3357	0.2692	0.2321	0.3506	0.3222	0.2564
3	shrestha14	0.2760	0.2062	0.2845	0.2308	0.2500	0.3052	0.4333	0.2223
4	weren14	0.2349	0.1914	0.2792	0.2949	0.1786	0.2013	0.2778	0.2211
5	villenaroman14	0.2315	0.1905	0.1961	0.3077	0.2321	0.2078	0.2667	0.2199
6	marquardt14	0.1998	0.1428	0.2102	0.1282	0.2679	0.1948	0.3111	0.1437
7	baker14	0.1677	0.1277	0.1678	0.1282	0.2321	0.1688	0.2111	0.1382
8	baseline	0.1404	0.0930	0.1820	0.0897	0.0536	0.1494	0.2333	0.1821
9	mechti14	0.1067	0.1244	0.1060	0.0897	0.1786	0.0584	0.1444	0.0451
10	castillojuarez14	0.0946	0.1445	0.1254	0.1795	0.0893	-	-	0.1236
11	ashok14	0.0834	0.1318	-	0.1282	-	0.1948	-	0.1291

In Figure 1 the average and standard deviation of the distances between predicted and true classes per subcorpus is shown. The highest distance on average is produced for reviews with a value of 1.69. The lowest distances on average and standard deviation are produced for Twitter. The similarity in distances between the social media subcorpora and the Spanish blogs is noteworthy. The complete list of distances among participants for each subcorpus is shown in Appendix B.

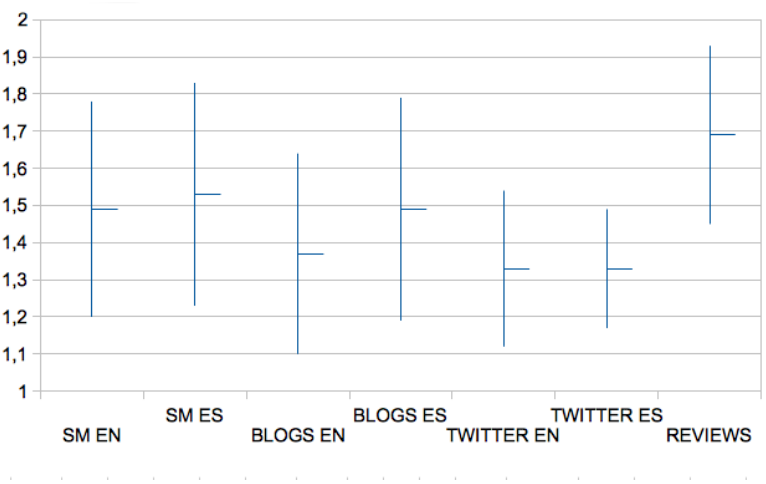

Figure 1. Distances between predicted and true classes per subcorpus.

In Appendix A, statistical significances of all pairwise system comparisons are detailed. As can be seen in Table A17, although lopezmonroy14 is the first in the general ranking, this system is statistically not significantly different from shrestha14, villenaroman14 and weren14. All systems are significantly different from the baseline, although weren, villenaroman and marquardt form a group close to baseline. It is noteworthy that

most of the systems are statistically indistinguishable regarding English social media, Spanish Twitter, and blogs (both languages).

With respect to age identification, all systems are significantly different from the baseline except ashok14 (the latter team did not participate in the Spanish task). There are some systems where differences are not statistically significant, such as lopezmonroy14 and liau14 or weren14 and villenaroman14. In blogs most of the systems are indistinguishable but significantly different from the baseline. On the other subcorpora, most of the systems are also different from the baseline. Looking at the accuracies the results show that most of the systems work significantly better than the baseline in age identification.

With respect to gender identification, all the systems are statistically different from the baseline, but lopezmonroy14, marquardt14, shrestha14, villenaroman14 and weren14 form a closer group. In English social media, English and Spanish blogs and Spanish Twitter, most of the systems are statistically not significantly different. Although all the systems are different from the baseline, most of them are statistically indistinguishable. Therefore, we cannot conclude that the systems perform better or worse than the baseline in gender identification. For example, in English social media all systems that are different from the baseline performed better in gender identification, in Twitter most of them performed better, but for Spanish social media the other way around happened and all the systems performed worse. The same happened in hotel reviews (in English) where most of the systems performed worse.

In Table 15 runtime results are shown. The fastest team was liau14 with bag-of-words features. With regard to the smallest data sets (Twitter and Blogs), we can make two groups depending on their runtime. The fastest teams utilised bag-of-words (liau14), words n-grams [18], style features [4], style and content features [20] or, in some cases, the second order features of [17]. In case of the largest subcorpora, such as social media and reviews, the difference among runtimes is more evident. The fastest ones also utilised simple content features and in some case stylistic ones. The slowest ones, with high difference, utilised IR-based features [32], parts-of-speech [13] or combinations of style and content-based features [19]. One of the slowest approaches [30] utilised term-vectors, but team participants reported that the low performance was due to the Weka library.

Table 15. Runtime performance (efficiency) per subcorpus.

Team	Twitter		Blogs		Social Media		Reviews
	EN	ES	EN	ES	EN	ES	EN
ashok14	3:23:36.00	-	5:57.22	-	18:26:49.00	19:03.24	
baker14	5:43.02	3:52.23	0:56.05	0:39.77	2:24:15.00	18:01.23	1:21.96
castillojuarez14	-	-	5:13.49	0:59.76	11:36:32.00	20:23.85	18:06.34
liau14	0:55.39	0:27.29	0:06.02	0:04.30	12:53.09	0:27.05	0:12.65
lopezmonroy14	7:02.91	5:36.05	3:47.04	3:22.02	34:06.53	6:25.89	4:01.40
marquardt14	1:47:15.00	35:06.63	not-known	7:36.18	36:05:51.00	2:08:14.00	5:44:45.00
mechti14	8:12.00	0:32.00	4:13.00	0:11.00	2:43:56.00	1:24.00	1:21:33.00
shrestha14	2:31.40	1:10.59	1:56.50	0:39.83	26:31.50	3:26.41	2:13.22
villenaroman14	1:12:22.00	38:28.70	10:06.74	8:04.18	69:55:12.00	9:14:15.00	5:38:07.00
weren14	41:32.38	1:33:48.00	4:46.46	4:06.79	30:18:02.00	2:34:33.00	1:17:29.00

We executed PAN-AP 2013 approaches for gender identification on the social media documents of PAN-AP 2014 (social media was the data used in PAN-AP 2013). A comparison for age identification was not possible due to the different age classes in PAN-AP 2013 and PAN-AP 2014. Most of the approaches failed at execution time so we only show those which could be executed. The only team with results for both years is lopezmonroy.[18] In Table 16 a comparison is shown. In English, although the best result was obtained by lopezmonroy13 [16], the majority of PAN-AP 2014 approaches obtained better results than PAN-AP 2013. In Spanish, results are more balanced between teams of the two years, although the two best results were obtained respectively by cagnina13 and haro13 [7]. The high number of approaches below the baseline in Spanish is noteworthy, as well as the higher accuracies obtained in Spanish than in English (being Spanish a gender-marked language). With respect to participants of both years, lopezmonroy13 achieved better results than lopezmonroy14 in English but not in Spanish.

Table 16. PAN-AP 2013 approaches evaluation results in PAN-AP 2014 social media in terms of accuracy on English (left) and Spanish (right) texts (gender identification).

English	
Team	Gender
lopezmonroy13	0.5438
villenaroman14	0.5421
liau14	0.5385
shrestha14	0.5382
weren14	0.5361
cagnina13	0.5287
lopezmonroy14	0.5237
marquardt14	0.5216
ashok14	0.5198
mecthi14	0.5198
baseline	0.5074
castillojuarez14	0.5053
haro13	0.5036
baker14	0.5012
ramirez13	0.4982
jimenez13	0.4967
patra13	0.4917

Spanish	
Team	Gender
cagnina13	0.6943
haro13	0.6855
liau14	0.6837
baseline	0.6555
shrestha14	0.6449
lopezmonroy14	0.6431
marquardt14	0.6431
lopezmonroy13	0.6336
weren14	0.6307
jimenez13	0.6237
mechti14	0.5919
villenaroman14	0.5724
ramirez13	0.5459
baker14	0.5000
castillojuarez14	0.4982

6 Conclusion

In this paper we present the results of the 2nd International Author Profiling Task at PAN-2014 within CLEF-2014. Given four different genres, namely, social media, blogs, Twitter, and hotel reviews, in the two languages English and Spanish, the 10 participants of the task had to identify gender and age of anonymous authors.

[18] lopezmonroy team was identified by pastor in PAN-AP 2013 (team obtaining the best performance)

The participants used several different features to approach the problem: content-based (bag of words, words n-grams, term vectors, named entities, dictionary words, slang words, contractions, sentiment words, and so on) and stylistic-based (frequencies, punctuations, POS, HTML use, readability measures and many different statistics). One participant [32] also combined many different IR-based features such as the cosine similarity or the Okapi BM25. This evaluation showed that good results were obtained by approaches which used simple content features (except the second order representation in [17] and the IR based features in [32]), for example bag-of-words (liau14), words n-grams [18] and term vectors [30]. Character n-grams demonstrated not to be a good approach for author profiling in general. The best results employed a second order representation based on relationships among terms, documents, profiles and subprofiles [17].

We draw following conclusions with respect to the different corpus parts: *a*) the highest accuracies were achieved on Twitter. We think this is due to the fact that we have a larger number of documents (tweets) per profile and the more spontaneous way to communicate in this social medium; *b*) the lowest results were obtained in English social media and hotel reviews, due to the lowest results in gender and age identification respectively; *c*) the highest distance between predicted and truth classes in age identification occurs in hotel reviews. A further analysis is needed in order to understand if for instance there are cases of deceptive opinions.

Acknowledgements The PAN task on author profiling has been organised in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie People Framework of the European Commission. We would like to thank Atribus by Corex for sponsoring the award for the winner team. We thank Julio Gonzalo, Jorge Carrillo and Damiano Spina from UNED for helping with the Twitter subcorpus. The work of the first author was partially funded by Autoritas Consulting SA and by Ministerio de Economía y Competitividad de España under grant ECOPORTUNITY IPT-2012-1220-430000 and CSO2013-43054-R. The work of the second author was in the framework the DIANA-APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

Bibliography

1. Yuridiana Aleman, Nahun Loya, Darnes Vilarino Ayala, and David Pinto. Two Methodologies Applied to the Author Profiling Task—Notebook for PAN at CLEF 2013. In Forner et al. [8].
2. Enrique Amigó, Jorge Carrillo-de-Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of RepLab 2014: author profiling and reputation dimensions for Online Reputation Management. In *Proceedings of the Fifth International Conference of the CLEF Initiative*, September 2014.
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. Christopher Ian Baker. Proof of Concept Framework for Prediction—Notebook for PAN at CLEF 2014. In Cappellato et al. [6].
5. 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.
6. 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.
7. Fermin Cruz, Rafa Haro, and Javier Ortega. ITALICA at PAN 2013: An Ensemble Learning Approach to Author Profiling—Notebook for PAN at CLEF 2013. In Forner et al. [8].
8. Pamela Forner, Roberto Navigli, and Dan Tufis, editors. *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, 2013.
9. 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. doi: http://dx.doi.org/10.1145/2348283.2348501.
10. 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. doi: http://doi.ieeecomputersociety.org/10.1109/DEXA.2012.55.
11. 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. doi: http://dx.doi.org/10.1007/978-3-642-40802-1_28.
12. Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. Stylometric analysis of bloggers' age and gender. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, *ICWSM*. The AAAI Press, 2009.
13. Gilad Gressel, Hrudya P, Surendran K, Thara S, Aravind A, and Prabaharan Poomachandran. Ensemble Learning Approach for Author Profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [6].
14. Janet Holmes and Miriam Meyerhoff. *The Handbook of Language and Gender*. Blackwell Handbooks in Linguistics. Wiley, 2003.

15. Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. literary and linguistic computing 17(4), 2002.
16. 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. [8].
17. 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. [6].
18. Suraj Maharjan, Prasha Shrestha, and Thamar Solorio. A Simple Approach to Author Profiling in MapReduce—Notebook for PAN at CLEF 2014. In Cappellato et al. [6].
19. James Marquardt, Golnoosh Fanardi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and Gender Identification in Social Media—Notebook for PAN at CLEF 2014. In Cappellato et al. [6].
20. Seifeddine Mechti, Maher Jaoua, and Lamia Hadrich Belguith. Machine learning for classifying authors of anonymous tweets, blogs and reviews—Notebook for PAN at CLEF 2014. In Cappellato et al. [6].
21. Michal Meina, Karolina Brodzinska, Bartosz Celmer, Maja Czokow, Martyna Patera, Jakub Pezacki, and Mateusz Wilk. Ensemble-based Classification for Author Profiling Using Various Features—Notebook for PAN at CLEF 2013. In Forner et al. [8].
22. Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 115–123, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
23. 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.
24. Eric W. Noreen. *Computer intensive methods for testing hypotheses: an introduction*. Wiley, New York, 1989.
25. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 37–44, New York, NY, USA, 2011. ACM.
26. James W. Pennebaker. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA, 2013.
27. 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.
28. Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstatios Stamatatos, and Giacommo Inches. Overview of the Author Profiling Task at PAN 2013—Notebook for PAN at CLEF 2013. In Forner et al. [8].

29. 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.
30. Julio Villena-Román and José-Carlos González-Cristóbal. DAEDALUS at PAN 2014: Guessing Tweet Author's Gender and Age—Notebook for PAN at CLEF 2014. In Cappellato et al. [6].
31. Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 783–792, 2010.
32. Edson R.D. Weren, Viviane P. Moreira, and José P.M. de Oliveira. Exploring Information Retrieval features for Author Profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [6].
33. 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.
34. 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

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

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

Table A1. Significance of accuracy differences between system pairs for age identification in the entire corpus.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	=
baker			=	***	***	=	***	***	***	***	***
castillojuarez				***	***	*	***	***	***	***	***
liau					=	***	***	=	***	**	***
lopezmonroy						***	***	=	*	=	***
marquardt							***	***	***	***	***
mechti								***	***	***	***
shrestha									**	=	***
villenaroman										=	***
weren											***
baseline											

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

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	***	***	***	**	*	***	***	***	***
baker			**	***	***	=	=	***	***	***	***
castillojuarez				***	***	=	***	***	***	***	***
liau					=	***	***	=	=	=	***
lopezmonroy						***	***	=	=	=	***
marquardt							***	***	***	***	***
mechti								***	***	***	***
shrestha									=	=	***
villenaroman										=	***
weren											***
baseline											

Table A3. Significance of accuracy differences between system pairs for age identification in Spanish social media.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			**	***	***	=	***	**	=	***	*
castillojuarez				***	***	***	=	***	***	***	=
liau					=	***	***	**	***	*	***
lopezmonroy						***	***	=	***	=	***
marquardt							***	***	=	***	*
mechti								***	***	***	**
shrestha									**	=	***
villenaroman										**	**
weren											***
baseline											

Table A4. Significance of accuracy differences between system pairs for age identification in English blogs.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	=	=	=	=	=	=	**	=
baker			=	=	=	=	=	=	=	=	**
castillojuarez				=	=	=	=	=	=	=	**
liau					=	=	=	=	=	=	**
lopezmonroy						=	*	=	=	=	***
marquardt							=	=	=	**	=
mechti								*	*	**	=
shrestha									=	=	**
villenaroman										=	**
weren											***
baseline											

Table A5. Significance of accuracy differences between system pairs for age identification in Spanish blogs.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	**
baker			=	=	=	=	=	=	=	*	**
castillojuarez				=	*	*	=	*	*	=	=
liau					=	=	=	=	=	*	**
lopezmonroy						=	=	=	=	*	**
marquardt							=	=	=	***	***
mechti								=	=	=	=
shrestha									=	*	***
villenaroman										*	**
weren											=
baseline											

Table A6. Significance of accuracy differences between system pairs for age identification in English Twitter.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	***	*	=	=	***	=	=	=	*
baker			***	**	**	=	***	=	=	=	=
castillojuarez				***	***	***	***	***	***	***	***
liau					=	*	***	=	*	**	***
lopezmonroy						**	***	=	=	**	***
marquardt							***	=	=	=	=
mechti								***	***	***	***
shrestha									=	*	***
villenaroman										=	**
weren											=
baseline											

Table A7. Significance of accuracy differences between system pairs for age identification in Spanish Twitter.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	=	***	***	***	***	***	***	***	***
baker			***	=	=	=	**	*	=	=	=
castillojuarez				***	***	***	***	***	***	***	***
liau					=	=	**	**	=	=	=
lopezmonroy						=	***	=	=	=	=
marquardt							**	=	=	=	=
mechti								***	**	***	**
shrestha									**	=	*
villenaroman										=	=
weren											=
baseline											

Table A8. Significance of accuracy differences between system pairs for age identification in English hotel reviews.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	***	***	=	***	***	***	***	*
baker			=	***	***	=	***	***	***	***	=
castillojuarez				***	***	=	***	***	***	***	*
liau					=	***	***	=	*	=	***
lopezmonroy						***	***	=	=	=	***
marquardt							***	***	***	***	*
mechti								***	***	***	***
shrestha									=	=	***
villenaroman										=	*
weren											***
baseline											

Table A9. Significance of accuracy differences between system pairs for gender identification in the entire corpus.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	**	***	***	***	***	***	***	***	***
baker			*	***	***	***	=	***	***	***	***
castillojuarez				***	***	***	***	***	***	***	***
liau					***	***	***	**	***	**	***
lopezmonroy						***	***	=	=	=	*
marquardt							***	***	***	***	*
mechti								***	***	***	***
shrestha									=	=	*
villenaroman										=	*
weren											*
baseline											

Table A10. Significance of accuracy differences between system pairs for gender identification in English social media.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	=	=	=	=	=	=	=	=
baker			=	**	=	=	=	**	**	**	=
castillojuarez				**	=	=	=	**	***	*	=
liau					=	=	=	=	=	=	*
lopezmonroy						=	=	=	*	=	=
marquardt							=	=	=	=	=
mechti								=	*	=	=
shrestha									=	=	*
villenaroman										=	**
weren											*
baseline											

Table A11. Significance of accuracy differences between system pairs for gender identification in Spanish social media.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			=	***	***	***	**	***	*	***	***
castillojuarez				***	***	***	**	***	=	***	***
liau					=	=	***	*	***	**	=
lopezmonroy						=	*	=	**	=	=
marquardt							*	=	**	=	=
mechti								*	=	=	*
shrestha									**	=	=
villenaroman										**	***
weren											=
baseline											

Table A12. Significance of accuracy differences between system pairs for gender identification in English blogs.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	**	**	=	=	=	*	**	=
baker			=	=	*	=	=	=	*	=	=
castillojuarez				=	*	=	=	=	=	=	=
liau					=	*	=	=	=	=	=
lopezmonroy						**	=	=	=	=	=
marquardt							=	=	*	=	=
mechti								=	=	=	=
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Table A13. Significance of accuracy differences between system pairs for gender identification in Spanish blogs.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			=	=	=	=	=	=	=	=	=
castillojuarez				=	=	=	=	=	=	=	=
liau					=	=	=	=	=	=	=
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								=	=	=	=
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Table A14. Significance of accuracy differences between system pairs for gender identification in English Twitter.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	***	***	**	=	=	**	=	=	**
baker			***	***	**	=	=	**	=	=	*
castillojuarez				***	***	***	***	***	***	***	***
liau					=	***	***	=	***	**	**
lopezmonroy						***	***	=	***	**	*
marquardt							=	*	=	=	=
mechti								*	=	=	=
shrestha									**	=	=
villenaroman										=	*
weren											=
baseline											

Table A15. Significance of accuracy differences between system pairs for gender identification in Spanish Twitter.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	=	***	***	***	***	***	***	***	***
baker			***	=	=	=	=	*	=	=	=
castillojuarez				***	***	***	***	***	***	***	***
liau					=	=	=	=	=	=	=
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								*	=	=	=
shrestha									=	=	**
villenaroman										=	=
weren											=
baseline											

Table A16. Significance of accuracy differences between system pairs for gender identification in English hotel reviews.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	***	***	***	=	***	***	***	***
baker			=	***	***	***	***	***	***	***	***
castillojuarez				***	***	*	=	***	***	***	***
liau					***	***	***	***	***	***	***
lopezmonroy						***	***	=	=	=	=
marquardt							***	***	***	***	***
mechti								***	***	***	***
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Table A17. Significance of accuracy differences between system pairs for joint identification in the entire corpus.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	**	***	***	***	***	***	***	***	***
baker			*	***	***	***	=	***	***	***	***
castillojuarez				***	***	***	***	***	***	***	***
liau					***	***	***	**	***	***	***
lopezmonroy						***	***	=	=	=	**
marquardt							***	***	***	***	*
mechti								***	***	***	***
shrestha									=	=	**
villenaroman										=	*
weren											*
baseline											

Table A18. Significance of accuracy differences between system pairs for joint identification in English social media.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	=	=	=	=	=	=	=	=
baker			=	**	=	=	=	**	**	**	=
castillojuarez				**	=	=	=	**	**	*	=
liau					=	=	=	=	=	=	**
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								=	*	=	=
shrestha									=	=	*
villenaroman										=	**
weren											*
baseline											

Table A19. Significance of accuracy differences between system pairs for joint identification in Spanish social media.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			=	***	***	***	**	***	=	***	***
castillojuarez				***	***	***	**	***	=	***	***
liau					=	=	***	=	***	**	=
lopezmonroy						=	*	=	**	=	=
marquardt							*	=	**	=	=
mechti								*	=	=	*
shrestha									**	=	=
villenaroman										**	***
weren											=
baseline											

Table A20. Significance of accuracy differences between system pairs for joint identification in English blogs.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	**	**	=	=	=	*	**	=
baker			=	=	*	=	=	=	*	=	=
castillojuarez				=	*	=	=	=	=	=	=
liau					=	*	=	=	=	=	=
lopezmonroy						**	=	=	=	=	=
marquardt							=	=	*	=	=
mechti								=	=	=	=
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Table A21. Significance of accuracy differences between system pairs for joint identification in Spanish blogs.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			=	=	=	=	=	=	=	=	=
castillojuarez				=	=	=	=	=	=	=	=
liau					=	=	=	=	=	=	=
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								=	=	=	=
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Table A22. Significance of accuracy differences between system pairs for joint identification in English Twitter.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	***	***	***	=	=	**	=	=	**
baker			***	***	***	=	=	**	=	=	**
castillojuarez				***	***	***	***	***	***	***	***
liau					=	**	**	=	***	**	*
lopezmonroy						***	**	=	**	**	*
marquardt							=	*	=	=	=
mechti								*	=	=	=
shrestha									**	=	=
villenaroman										=	**
weren											=
baseline											

Table A23. Significance of accuracy differences between system pairs for joint identification in Spanish Twitter.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	=	***	***	***	***	***	***	***	***
baker			***	=	=	=	=	*	=	=	=
castillojuarez				***	***	***	***	***	***	***	***
liau					=	=	=	=	=	=	=
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								*	=	=	=
shrestha									=	=	*
villenaroman										=	=
weren											=
baseline											

Table A24. Significance of accuracy differences between system pairs for joint identification in English hotel reviews.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	***	***	***	=	***	***	***	***
baker			=	***	***	***	***	***	***	***	***
castillojuarez				***	***	**	=	***	***	***	***
liau					***	***	***	***	**	***	***
lopezmonroy						***	***	=	=	=	=
marquardt							***	***	***	***	***
mechti								***	***	***	***
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Appendix B Distances in Age Identification

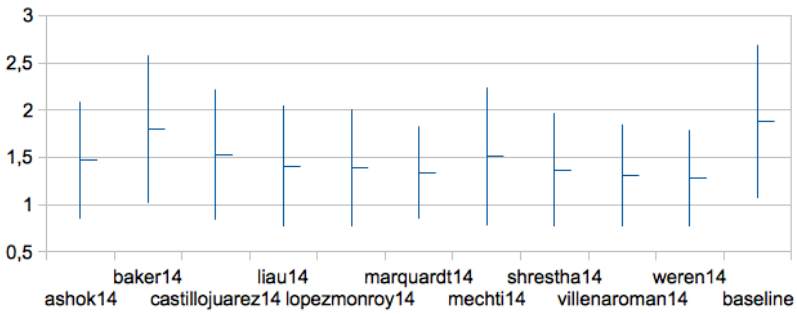

Figure B1. Distances between predicted and truth classes in English social media.

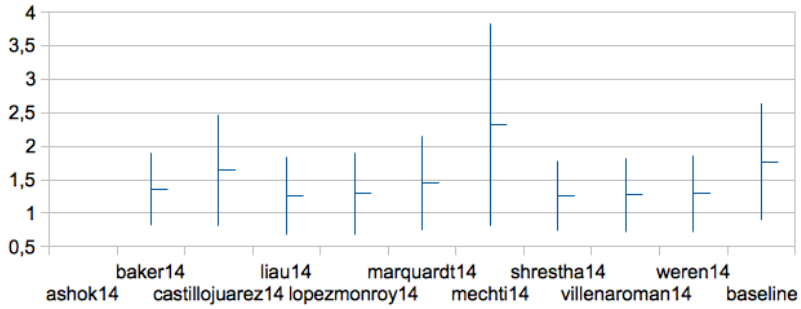

Figure B2. Distances between predicted and truth classes in Spanish social media.

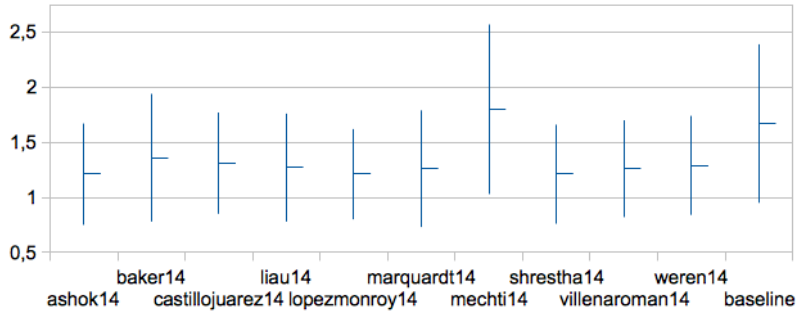

Figure B3. Distances between predicted and truth classes in English blogs.

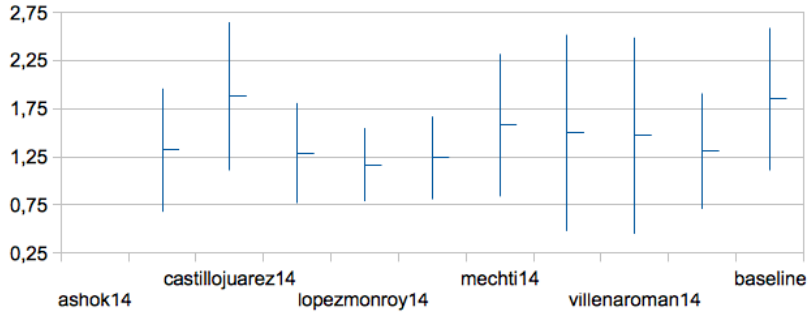

Figure B4. Distances between predicted and truth classes in Spanish blogs.

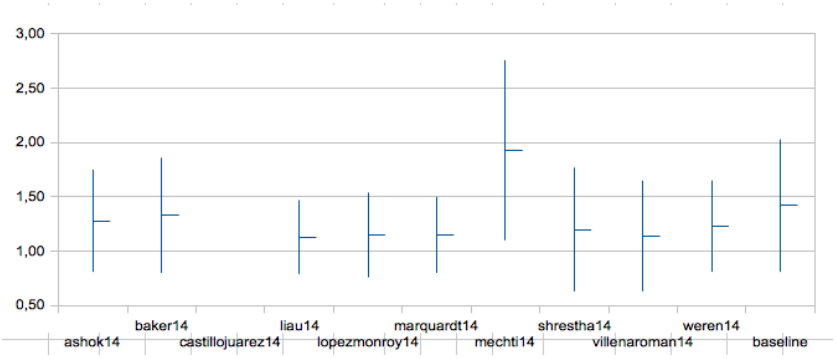

Figure B5. Distances between predicted and truth classes in English Twitter.

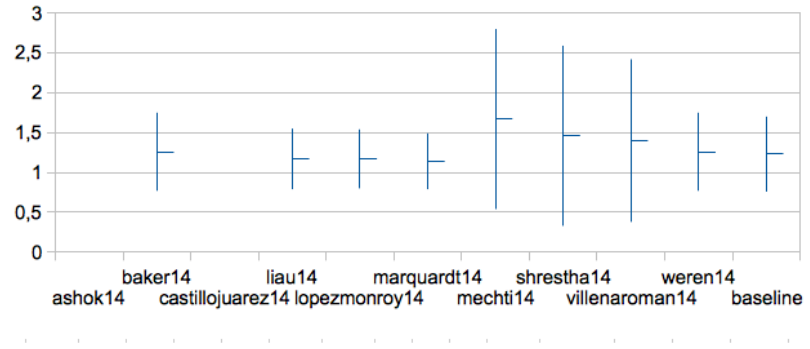

Figure B6. Distances between predicted and truth classes in Spanish Twitter.

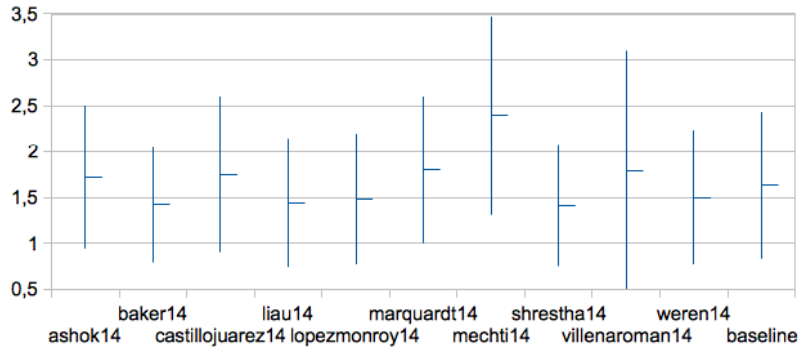

Figure B7. Distances between predicted and truth classes in English hotel reviews.