Overview of the Author Profiling Task at PAN 2013

Francisco Rangel,[1 2] Paolo Rosso,[2] Moshe Koppel,[3] Efstathios Stamatatos,[4]
Giacomo Inches[5]

[1]Autoritas Consulting, S.A., Spain
[2]Natural Language Engineering Lab, ELiRF, Universitat Politècnica de València, Spain
[3]Dept. of Computer Science, Bar-Illan University, Israel
[4]Dept. of Information and Communication Systems Engineering, University of the Aegean, Greece
[5]Information Retrieval Group, Faculty of Informatics, University of Lugano, Switzerland

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

Abstract This overview presents the framework and results for the Author Profiling task at PAN 2013. We describe in detail the corpus and its characteristics, and the evaluation framework we used to measure the participants performance to solve the problem of identifying age and gender from anonymous texts. Finally, the approaches of the 21 participants and their results are described.

1 Introduction

In classical authorship attribution, we are given a closed set of candidate authors and are asked to identify which one of them is the author of an anonymous text. Author profiling, on the other hand, distinguishes between classes of authors, rather than individual authors. Thus, for example, profiling is used to determine an author's gender, age, native language, personality type, etc. Author profiling is a problem of growing importance in a variety of areas, including forensics, security and marketing. For instance, from a forensic linguistics perspective, being able to determine the linguistic profile of the author of a suspicious text solely by analyzing the text could be extremely valuable for evaluating suspects. Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, what types of people like or dislike their products. Here we consider the problem of author profiling in social media, with particular focus on the use of everyday language and how this reflects basic social and personality processes. Our starting point is the seminal work of Argamon *et al.* [3], where it was shown that statistical analysis of word usage in documents could be used to determine an author's gender, age, native language and personality type.

In PAN 2013[1] we consider the gender and age aspects of the author profiling problem, both in English and Spanish. So far research work in computational linguistics [2] and social psychology [26] has been carried out mainly for English. We believe it is interesting to investigate gender and age classification task in a language other than

[1] http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/index.html

English, therefore considering Spanish, too.

In Section 2 we present the state of the art, describing related work and how the task has been approached. In Section 3 we describe the details of the collection used and the evaluation measures. In Section 4 we present the authors' approaches and we discuss the results in Section 5, concluding the overview in Section 6.

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, natural language processing. 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.* [2] analyzed 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[13], Burger and Henderson[4]) have also investigated obtaining 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.* [16] 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.* [30] studied the effect of age and gender in the style of writing 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 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 detection and to 89.2% in gender detection.

It is to be noted that the previously described studies were conducted with texts of at least of 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 [33] 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.* [22] 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 [21]), and used an approach based on logistic regression. They also measured the effect of the gender in the performance of age

detection, 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 the author profiling problem is the need to obtain labelled data for the authors, for example, to obtain their age and gender. Studies in classical literature deals 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 [22] 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[2],where authors report their gender and exact age, and Koppel *et al.* [16], who retrieved the dataset from Blogspot[3]. In these cases we have to be aware of a common issue, the use of these media (mainly blogs) to promote web positions in search engines through the use of false profiles. This is likely to introduce noise to the evaluation corpus, but it also reflects the realistic state of the available data.

3 Evaluation framework

In this section we describe the data collection obtained for the task, its properties, challenges and novelties as well the evaluation measures.

3.1 Data collection

We built the corpus with thousands of blog posts taking into account that:

- The variety of themes provides a wide spectrum of topics, making the task of determining age and gender more realistic. The ample diversity of topics allows to investigate standard cliches, for example, men speaking a lot about beer or football and women about nails or shopping, for breaking or reinforcing them.
- Blog posts are used daily for search engine optimization and can be automatically generated by robots or be advertisements (chatbots).
- People may use social media to talk also about sex and few can also break the line and use these systems to misbehave and engage in conversations that may result into sexual harassment. For this reason and due to the importance of unveiling fake profiles, we decided to test the robustness of the author profiling approaches including in our collection some texts from last year PAN task on sexual predator identification.
- We wanted to carry out the task in a multilingual setting, therefore, in addition to English we included a Spanish part in our collection. Spanish and English are two of the most used languages in the world[4].

[2] http://www.netlog.com
[3] http://blogspot.com
[4] http://www.internetworldstats.com/stats7.htm

We looked online for open and public repositories such as Netlog with posts labelled with author demographics such as gender and age. Once found, we decided to group posts by author, selecting those authors with at least one post, and chunking in different files those authors with more than 1,000 words in their posts. We also included authors with very few and possibly short posts in order to maintain a realistic evaluation framework. We divided the collection into the following parts: training, early bird evaluation and final testing. Authors were randomly split into these parts, making sure that each author is included in exactly one part. For age detection, we followed what was previously done in [30] and considered three classes: 10s (13-17), 20s (23-27) and 30s (33-47). The collection is balanced by gender and imbalanced by age group. Additionally, trying to preserve a real-world scenario[5] , we incorporated a small number of samples from conversations of sexual predators [14] together with samples from adult-adult conversations about sex. In Table 1 we illustrate the statistics of English and Spanish collections.

Table 1. Corpus statistics for training, early bird evaluation and test.

Lang	**Age**	**Gender**	**No. of Authors**		
			Training	Early Bird	Test
en	10s	male	8 600	740	888
		female	8 600	740	888
	20s	male	(72) 42 828	3 840	(32) 4 576
		female	(25) 42 875	3 840	(10) 4 598
	30s	male	(92) 66 708	6 020	(40) 7 184
		female	66 800	6 020	7 224
Σ			236 600	21 200	25 440

Lang	**Age**	**Gender**	**No. of Authors**		
			Training	Early Bird	Test
es	10s	male	1 250	120	144
		female	1 250	120	144
	20s	male	21 300	1 920	2 304
		female	21 300	1 920	2 304
	30s	male	15 400	1 360	1 632
		female	15 400	1 360	1 632
Σ			75 900	6 800	8 160

In the training part of the English collection, numbers inside parentheses for male 20s and 30s correspond to the number of samples of sexual predator conversations

[5] E.g. There are statistics of about 200 tweets per hour in English from sexual predators (http://www.mirror.co.uk/news/uk-news/paedophiles-using-twitter-to-find-victims-1253833). Twitter issued about 200 million tweets per day (https://blog.twitter.com/2011/200-million-tweets-day) in 2011, achieving 400 million tweets per day in 2013 (http://www.webpronews.com/twitter-turns-7-boasts-400m-tweets-per-day-2013-03). This is about 0.0012%

while numbers inside parenthesis for female 20s correspond to the adult-adult sexual conversation samples. We provided these samples for training purposes. In the collection for early bird evaluation, we did not include any sample of this kind. The final collection was built adding a 20% of samples over the early bird dataset this time including samples from sexual predator conversations for male 20s and 30s, and samples from adult-adult conversations for female 20s.

The distribution of number of words per document for each language is depicted in Figure 3.1.

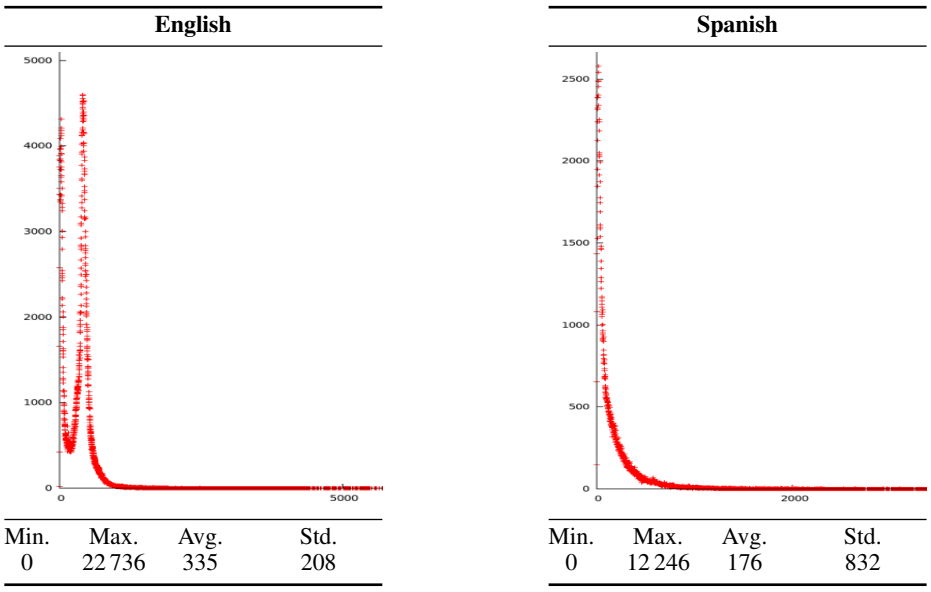

Figure 1. Distribution of the number of words per document

As can be seen, there are significant differences between the two languages. More than 80% of Spanish posts are about 15-word long (e.g. greetings, especially for teenagers). On the other hand, English speakers seem to describe situations, experiences or thoughts, but in a more elaborated way.

3.2 Performance measures

For evaluating participants' approaches we have used accuracy. Concretely, we calculated the ratio between the number of authors correctly predicted by the total number of authors. We calculated separately the accuracy for each language, gender, and age group. 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 language.

We also calculated the total number of correctly identified gender and age for predator samples, in order to determine what approaches are more robust to this kind of outliers. Finally, we calculated the total time needed to process the test data, in order to investigate the difficulties of processing big volumes of data in the framework of a real-world application.

4 Overview of the participants' approaches

We received 21 submissions for the task of Author Profiling and a total of 18 notebook papers: 8 long papers and 10 short papers. We present the analysis of the 18 approaches we received a description of.

Pre-processing . Only few participants preprocessed the data. Various participants [23][20][19][32][24] cleaned HTML to obtain plain text, one participant [9] deleted those documents containing at least 0.1% of spam words and another participant [17] used Principal Component Analysis to linearly reduce the dimensionality. During the training phase, some participants [7][9][20][8][29] selected a subset from the training data in order to reduce dimensionality. Only one participant [19] tried to discriminate between human-like posts and spam-like posts or chatbots.

Features . Many participants [17] [5] [24] [23] [6] [19] [9] [1] [28] used stylistic features such as frequencies of punctuation marks, capital letters, quotations, and so on, together with POS tags [17] [19] [1] [5] [28] or HTML-based features as image urls or links [28] [29] [19]. Readability features has been widely used in several approaches [23] [17] [19] [9] [1] [32] [11]. In the last approach readability features were the only ones used. Emoticons were used by two participants [1] [8] and discarded from one participant [29].

Different content features (e.g. Latent Semantic Analysis, bag of words, TF-IDF, dictionary-based words, topic-based words, entropy-based words, and so on) were also used by many participants [29] [23] [17] [31] [7] [9] [19] [5] [28] [24] [8]. Different participants considered named entities [9], sentiment words [23], emotion words [19], [9], [8], and slang, contractions and words with character flooding [9] [7] [1] [8].

A different approach based on information retrieval was presented by one participant [32]. In such approach, the text to be identified was used as a query for a search engine. One participant [6] introduced a high variety of corpus statistics to build unsupervised features and four participants [19] [15] [20] [29] used n-grams models. Finally, one participant [19] introduced advanced linguistic features such as collocations and another participant [18] used second order representation based on

relationships between documents and profiles.

Classification approaches . All the approaches used supervised machine learning methods. The vast majority of them [28] [23] [31] [11] [32] used decision trees. Three approaches [17] [5] [29] used Support Vector Machines, two approaches [6] [9] used logistic regression, and the rest used Naïve Bayes [19], Maximum Entropy [24], Stochastic Gradient Descent [7] and random forest [1].

5 Evaluation of the participants' approaches and discussion

We divided the evaluation in two steps, an early bird option for those who wanted to test their approaches before the final submission in order to have some feedback, and the final evaluation. There were 5 early bird submissions and 21 for final evaluation. We could not evaluate one early bird submission due to runtime errors on the TIRA[6] platform. A baseline was provided in order to compare the different approaches with. This baseline was programmed as two random classifiers for each variable (gender and age group), obtaining 50% of accuracy for gender identification and 33% for age identification, and 16.5% for joint identification.

In Table 2 the performance of early bird submissions is shown. In Table 3 the final ranking for each language is presented. We show the accuracy for gender and age group and the accuracy for the joint identification. The difficulty of the task is reflected in the low values of such measure, especially for gender identification with close to the baseline. In addition, the joint identification shows a dramatic decrease in the result, highlighting the even greater difficulty of the joint identification.

Table 2. Evaluation results for early birds in terms of accuracy on English (left) and Spanish (right) texts.

English			
Team	Total	Gender	Age
Ladra	0.3301	0.5631	0.5924
Gillam	0.3245	0.5413	0.5947
Jankowska	0.2796	0.5185	0.5463
baseline	0.1649	0.4997	0.3324
Aleman	0.0162	0.0277	0.0278

Spanish			
Team	Total	Gender	Age
Ladra	0.3541	0.6171	05757
Jankowska	0.2724	0.5834	0.4479
Gillam	0.2521	0.4774	0.5357
baseline	0.1653	0.5001	0.3353
Aleman	0.0490	0.0844	0.0841

In order to determine the overall performance, we calculated the average of the total values for English and Spanish. The [18] team obtained the overall best performance on average in English and Spanish.

[6] http://tira.webis.de

Table 3. Evaluation results in terms of accuracy on English (left) and Spansih (right) texts.

English			
Team	Total	Gender	Age
Meina	0.3894	0.5921	0.6491
Pastor L.	0.3813	0.5690	0.6572
Seifeddine	0.3677	0.5816	0.5897
Santosh	0.3508	0.5652	0.6408
Yong Lim	0.3488	0.5671	0.6098
Ladra	0.3420	0.5608	0.6118
Aleman	0.3292	0.5522	0.5923
Gillam	0.3268	0.5410	0.6031
Kern	0.3115	0.5267	0.5690
Cruz	0.3114	0.5456	0.5966
Pavan	0.2843	0.5000	0.6055
Caurcel Diaz	0.2840	0.5000	0.5679
H. Farias	0.2816	0.5671	0.5061
Jankowska	0.2814	0.5381	0.4738
Flekova	0.2785	0.5343	0.5287
Weren	0.2564	0.5044	0.5099
Sapkota	0.2471	0.4781	0.5415
De-Arteaga	0.2450	0.4998	0.4885
Moreau	0.2395	0.4941	0.4824
baseline	0.1650	0.5000	0.3333
Gopal Patra	0.1574	0.5683	0.2895
Cagnina	0.0741	0.5040	0.1234

Spanish			
Team	Total	Gender	Age
Santosh	0.4208	0.6473	0.6430
Pastor L.	0.4158	0.6299	0.6558
Cruz	0.3897	0.6165	0.6219
Flekova	0.3683	0.6103	0.5966
Ladra	0.3523	0.6138	0.5727
De-Arteaga	0.3145	0.5627	0.5429
Kern	0.3134	0.5706	0.5375
Yong Lim	0.3120	0.5468	0.5705
Sapkota	0.2934	0.5116	0.5651
Pavan	0.2824	0.5000	0.5643
Jankowska	0.2592	0.5846	0.4276
Meina	0.2549	0.5287	0.4930
Gillam	0.2543	0.4784	0.5377
Moreau	0.2539	0.4967	0.5049
Weren	0.2463	0.5362	0.4615
Cagnina	0.2339	0.5516	0.4148
Caurcel Diaz	0.2000	0.5000	0.4000
H. Farias	0.1757	0.4982	0.3554
baseline	0.1650	0.5000	0.3333
Aleman	0.1638	0.5526	0.2915
Seifeddine	0.0287	0.5455	0.0512
Gopal Patra	–	–	–

It is difficult to establish a correlation between the used features in the different approaches and the obtained results, due mainly to the amount of shared features in all of them. It is to be noted the usage of second order representations based on relationships between documents and profiles by the winner of the task [18] and the use of collocations for the winner of the English task [19], features that do not seem to be as good for Spanish (or maybe more difficult to tune). Stylistic and content features were used for the vast majority of approaches and the obtained values for accuracy show results in different positions of the ranking. POS features were used in five different approaches, e.g. by systems in the first position for English [19] and in the first position for the Spanish [28], with values under the median of the ranking for the rest of the approaches. Such features seem to improve the performance on the task. Readability is another feature widely used for the vast majority of the approaches. We can compare the performance of this feature with the rest because there is an approach [11] based only on such feature, achieving the 8th position in English and the 13th in Spanish. Except one approach [19], those which used n-gram features did not achieve very good results, all of them over the median of the ranking. The use of sentiment words [23] and emotion words [9] [8] does not seem to improve the accuracy, in the same manner than the use of slang words [9] [7] [1] [8], although these approaches used many other features and it is difficult to establish a correlation.

Regarding employing some kind of preprocessing, it is interesting that except two cases [19] [17] the rest get worse performance, although it may be probably due to the

features used not to the preprocessing itself.

Table 4. Number (and accuracy) of adult-adult sexual conversations (left) and predators (right) correctly identified.

Team	**Adult-Adult**			**Predators**		
	Total	Gender	Age	Total	Gender	Age
Aleman	1 (0.1)	3 (0.3)	2 (0.2)	26 (0.36)	53 (0.74)	34 (0.47)
Cagnina	4 (0.4)	4 (0.4)	7 (0.7)	8 (0.11)	24 (0.33)	9 (0.13)
Caurcel Diaz	0 (0.0)	0 (0.0)	0 (0.0)	40 (0.56)	72 (1.00)	40 (0.56)
Cruz	0 (0.0)	0 (0.0)	8 (0.8)	41 (0.57)	69 (0.96)	44 (0.61)
De Arteaga	1 (0.1)	6 (0.6)	2 (0.2)	14 (0.19)	27 (0.38)	31 (0.43)
Flekova	4 (0.4)	4 (0.4)	4 (0.4)	34 (0.47)	61 (0.85)	39 (0.54)
Gillam	0 (0.0)	1 (0.1)	4 (0.4)	30 (0.42)	72 (1.00)	30 (0.42)
Gopal Patra	1 (0.1)	5 (0.5)	4 (0.4)	12 (0.17)	55 (0.76)	17 (0.24)
H. Farias	1 (0.1)	4 (0.4)	2 (0.2)	26 (0.36)	55 (0.76)	34 (0.47)
Jankowska	0 (0.0)	1 (0.1)	0 (0.0)	44 (0.61)	72 (1.00)	44 (0.61)
Kern	9 (0.9)	9 (0.9)	9 (0.9)	25 (0.35)	47 (0.65)	35 (0.49)
Ladra	9 (0.9)	9 (0.9)	9 (0.9)	33 (0.46)	72 (1.00)	33 (0.46)
Meina	6 (0.6)	6 (0.6)	8 (0.8)	41 (0.57)	72 (1.00)	41 (0.57)
Moreau	2 (0.2)	4 (0.4)	4 (0.4)	19 (0.26)	33 (0.46)	39 (0.54)
Pastor L.	0 (0.0)	1 (0.1)	8 (0.8)	32 (0.44)	72 (1.00)	32 (0.44)
Pavan	0 (0.0)	0(0.0)	0 (0.0)	50 (0.56)	72 (1.00)	40 (0.56)
Santosh	9 (0.9)	9 (0.9)	9 (0.9)	29 (0.40)	69 (0.96)	32 (0.44)
Sapkota	0 (0.0)	9 (0.9)	0 (0.0)	9 (0.13)	12 (0.17)	40 (0.56)
Seifeddine	2 (0.2)	2 (0.2)	6 (0.6)	20 (0.28)	52 (0.72)	29 (0.40)
Weren	0 (0.0)	1 (0.1)	0 (0.0)	39 (0.54)	71 (0.99)	40 (0.56)
Yong Lim	1 (0.1)	6 (0.6)	1 (0.1)	17 (0.24)	28 (0.39)	30 (0.42)

In Table 4 the identification of fake profiles for sexual predators is shown. The first group of columns shows the number of correctly identified profiles for adult-adult sexual conversations and the second group shows the number of correctly identified fake profiles for sexual predators. In brackets the ratio is shown.

The vast majority of participants identified correctly cases of adult-adult sexual conversations but what is more surprising is that all the participants identified the right age and gender of many predator samples. At least 7 participants identified more than 50% of such cases, 10 participants identified gender for more than 95% of the cases and 7 participants identified age for more than 50% of them. Best results were obtained by 3 participants who combined content and stylistic features [5] [19] [24], one participant who used n-grams [15] and one participant who used a content-based approach improved with specific dictionaries (slang, contractions...) [7]. The approach based on Information Retrieval techniques [32] also obtained top results. The approach based only on the readability features [11] obtained 42% of accuracy, meaning that such features have an important impact on detecting such cases.

Table 5. Runtime performance in milliseconds, and in minutes, hours or days.

Team	Runtime	
Gillam	615 347ms	10.26min
Ladra	1 729 618ms	28.83min
Pastor L.	2 298 561ms	38.31min
Caurcel Diaz	3 241 899ms	54.03min
Pavan	3 734 665ms	1.04h
De Arteaga	3 940 310ms	1.09h
Cruz	9 559 554ms	2.66h
Weren	11 684 955ms	3.25h
Jankowska	16 761 536ms	4.66h
Santosh	17 511 633ms	4.86h
Kern	18 285 830ms	5.08h
Flekova	18 476 373ms	5.13h
Gopal Patra	22 914 419ms	6.37h
Aleman	23 612 726ms	6,56h
H. Farias	24 558 035ms	6.82h
Sapkota	64 350 734ms	17.88h
Meina	383 821 541ms	4.44d
Moreau	448 406 705ms	5.19d
Yong Lim	577 144 695ms	6.68d
Cagnina	855 252 000ms	9.90d
Seifeddine	1 018 000 000ms	11.78d
-	-	

Finally, in Table 5 we show the time each participant needed to finish the task, reversely ordered by runtime. Runtime is shown in milliseconds. The differences between the fastest (10.26 minutes) [11] and the slowest (11.78 days) [31] is enormous. The fastest [11] approached the task only with the readability features, obtaining the 8th position in English and 13th in Spanish. The slowest [31] approached the task with content features, obtaining the 3rd. position in English and 21th in Spanish. The vast majority of approaches took a few hours. The slowest participants used collocations [19], POS [17], n-grams [20] and performed preprocessing such as html removal [20] [19], detection of chatbots [19] and Principal Component Analysis [17].

6 Conclusions

In this paper we present the results of the 1st International Author Profiling Task at PAN-2013 within CLEF-2013. Given a large and realistic collection of blog posts and chat logs, the 21 participants of the task had to identify gender and age of anonymous authors.

Participants used several different features to approach the problem, being able to be grouped into content-based (bag of words, named entities, dictionary words, slang words, contractions, sentiment words, emotion words, and so on), stylistic-based (frequencies, punctuations, POS, HTML use, readability measures and many different statistics), n-grams based, IR-based and collocations-based. Results show the difficulty of the task, mainly for the gender identification and for the joint identification of gender and age.

We introduced some conversations from sexual predators in order to check the robustness of the approaches, and we were pleasantly surprised by the high amount of such cases correctly identified by all the participants.

Acknowledgements

The author profiling task @PAN-2013 was an activity of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie People Framework of the European Commission. We want to thank the Forensic Lab of the Universitat Pompeu Fabra Barcelona for sponsoring the award for the winner team. 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. 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. The work of fifth author was funded in part by the Swiss National Science Foundation (SNF) project "Mining Conversational Content for Topic Modelling and Author Identification (ChatMiner)" under grant number 200021_130208.

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. [10].

[2] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346, 2003.

[3] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2): 119–123, February 2009.

[4] 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.

[5] 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. [10].

[6] Maria De-Arteaga, Sergio Jimenez, George Duenas, Sergio Mancera, and Julia Baquero. Author Profiling Using Corpus Statistics, Lexicons and Stylistic Features—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[7] Andres Alfonso Caurcel Diaz and Jose Maria Gomez Hidalgo. Experiments with SMS Translation and Stochastic Gradient Descent in Spanish Text Author Profiling—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[8] Delia Irazu Hernandez Farias, Rafael Guzman-Cabrera, Antonio Reyes, and Martha Alicia Rocha. Semantic-based Features for Author Profiling Identification: First insights—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[9] Lucie Flekova and Iryna Gurevych. Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[10] Pamela Forner, Roberto Navigli, and Dan Tufis, editors. *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, 2013.

[11] Lee Gillam. Readability for author profiling?—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[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] Janet Holmes and Miriam Meyerhoff. *The Handbook of Language and Gender*. Blackwell Handbooks in Linguistics. Wiley, 2003. ISBN 9780631225027.

[14] Giacomo Inches and Fabio Crestani. Overview of the International Sexual Predator Identification Competition at PAN-2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy*, September 2012.

[15] Magdalena Jankowska, Vlado Keselj, and Evangelos Milios. CNG Text Classification for Authorship Profiling Task—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[16] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender, 2003.

[17] Wee Yong Lim, Jonathan Goh, and Vrizlynn L. L. Thing. Content-Centric Age and Gender Profiling—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[18] 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. [10].

[19] 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. [10].

[20] Erwan Moreau and Carl Vogel. Style-based Distance Features for Author Profiling—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[21] 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.

[22] 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.

[23] Braja Gopal Patra, Somnath Banerjee, Dipankar Das, Tanik Saikh, and Sivaji Bandyopadhyay. Automatic Author Profiling Based on Linguistic and Stylistic Features—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[24] Aditya Pavan, Aditya Mogadala, and Vasudeva Varma. Author Profiling Using LDA and Maximum Entropy—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[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. ISBN 9781608194964.

[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] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. Author Profiling: Predicting Age and Gender from Blogs—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[29] Upendra Sapkota, Thamar Solorio, Manuel Montes-Y-Gomez, and Gabriela Ramirez-De-La-Rosa. Author Profiling for English and Spanish Text—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[30] 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.

[31] Mechti Seifeddine, Jaoua Maher, and Hadrich Belghith Lamia. Author Profiling Using Style-based Features—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[32] Edson Weren, Viviane P. Moreira, and Jose Oliveira. Using Simple Content Features for the Author Profiling Task—Notebook for PAN at CLEF 2013. In Forner et al. [10].

[33] Cathy Zhang and Pengyu Zhang. Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA, 2010.