I, me, mine: the Role of Personal Phrases in Author Profiling

Rosa María Ortega-Mendoza[1,2], Anilú Franco-Arcega[1], Adrián Pastor López-Monroy[3], and Manuel Montes-y-Gómez[3]

[1] Universidad Autónoma del Estado de Hidalgo (UAEH), Mexico.
{or300944,afranco}@uaeh.edu.mx
[2] Instituto Tecnológico Superior del Oriente del Estado de Hidalgo (ITESA), Mexico.
mortega@itesa.edu.mx
[3] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico.
{pastor,mmontesg}@inaoep.mx

Abstract. The Author Profiling (AP) task aims to distinguish between groups of authors labeled by a common demographic characteristic such as gender or age by studying the language usage. In this work we studied the role of personal phrases (i.e., sentences containing first person pronouns) for the AP task. We support the idea that people better expose their personal interests and writing style when they talk about themselves and, consequently, that words near to a personal pronoun reveal valuable information for the classification of authors. The evaluation using different social media data showed that phrases containing *singular first person pronouns* are highly valuable for predicting the age and gender of users. Considering only these phrases we obtained reductions of up to 60% of the information in the user documents and a comparable classification performance than using all available data. In addition, the results obtained by personal phrases considerably outperformed those from non-personal sentences, indicating their greater suitability for the AP task. We consider these findings could be further applied in the design of strategies for the construction of AP corpora, novel feature selection methods, as well as new feature and instance weighting schemes.

Keywords: Author profiling, personal pronouns, topics, writing style.

1 Introduction

In Natural Language Processing, the Author Profiling (AP) task consists in analyzing texts in order to extract as much information as possible from their authors [11]. Its aim is to predict general or demographic attributes that integrate authors' profiles such as: gender [11, 2, 31, 12], age [2, 31, 20, 12], personality [1, 32], native language [2], political orientation [21], among others. Recently, because of the variety of its applications, AP has gained a lot of interest. For example, in marketing, companies leverage online reviews to improve targeted advertising, and in forensics, the linguistic profile of authors could be used as valuable additional evidence.

AP is supported on the idea that documents are the major medium by which people communicate their knowledge and express their thoughts and opinions. It also considers that word usage patterns extracted from these documents expose people interests and writing style, which in turn, could reveal valuable information for their automatic profiling. Broadly speaking, AP has been approached as a single-label classification problem using machine learning algorithms [33]. In this context, most of the work has been devoted to determine useful textual features to model the writing profile of authors [1, 11, 28, 31]. According to the literature two kinds of features are the most relevant: thematic features, mainly captured by nouns, verbs and adjectives, and stylistic features, e.g., function words, punctuation marks, and POS tags [14].

In this work, rather than define a suitable set of features for AP, we focus on studying the relevance of sentences containing first person pronouns, which we refer as *personal phrases*. Our interest in this kind of phrases is motivated by recent works in social psychology, which have demonstrated that pronouns and prepositions reveal important information about the linguistic profile of an author [22], and also people tend to be more honest when they write about themselves [19]. Based on these findings, we hypothesize that words around personal pronouns better expose the thematic interests and writing style of authors, and therefore that they could reveal valuable information for their classification. Accordingly, the research questions we aim to answer are:

- Are all the information in a document equally relevant for AP? Particularly, are personal phrases more discriminating than others?
- Are the personal phrases containing singular and plural first person pronouns equally useful for AP? Are they complementary or redundant?
- Do personal phrases better expose the writing style or the thematic interests of authors?
- Are personal phrases equally relevant in different social media domains?

To answer these questions we evaluated the prediction of users' age and gender in different social media domains. Our study shows that personal phrases can be considered the *essence of documents*[4] for the AP task [16]. We mainly found that focusing on the subset of personal phrases, it is possible to get reductions of up to 60% of the information in the user documents, while maintaining the classification performance. Our findings have significant implications for future work in AP, since they can lead to the design of new feature selection and weighting methods as well as to the development of alternative strategies for the construction of AP corpora.

The rest of the paper is organized as follows. Section 2 describes some previous works in AP, making special emphasis on psychological motivated approaches. Section 3 presents the corpora used in the experiments, whereas Section 4 describes our experimental methodology. Section 5 presents the experi-

[4] In this context, documents are commonly referred to as *user profiles* or *user histories*, and they correspond to all textual information generated by a user, for example, all posts from her blog or the set of tweets from her account.

ments and results in different social media. Finally, Section 6 depicts our conclusions and some future work directions.

2 Related work

There are several works for AP in social media [31, 21, 32]. These works have mainly proposed different document representations, which combine several kinds of features [24]. For example, Argamon et al. [2] used content and style features to identify the age, gender, native language and neuroticism level of authors. Mukherjee & Liu [17] studied the classification of blogs by gender using POS patterns as features. Other proposals include the use of stylometrics characteristics. For example, Goswami et al. [9] predicted age and gender of blogs' authors by means of slang words and the length of sentences. Rangel & Rosso [25] used style features such as the frequency of capital letters, words length, and number of words with flooded characters (e.g. Heeeellooo). Meina et al. [15] have studied structural features such as the number of sentences, words, paragraphs, special characters, among others. On the other hand, there are some works that have also explored the use of sociolinguistic features to determine the age and gender of authors [29]. This kind of features aims to capture, for example, the communication behavior (e.g. retweet frequency) and the network characteristics (e.g. number of followers and friends) of social media users.

From a psychological perspective, some recent works have shown that language carries information about our feelings, emotions [27, 26], and opinions [34], and that function words are the most revealing [4, 22]. For example, the frequent use of singular first person pronouns is related to: young people [23], female [18, 2], low social status [10], and depression [30]. Furthermore, it has been found that people tend to use this kind of pronouns when they tell the truth [19]. In other words, the use of self-references such as "I", "me", "my" and "mine" are strongly related to the expression of people's feelings, concerns and opinions.

These previous works have demonstrated the usefulness of pronouns as features for characterizing the author of a document. This paper goes a step forward by studying the role of personal phrases in AP across different social media domains. We consider that words around personal pronouns better expose the thematic interests and writing style of social media users, and that this subset of phrases could be considered as the essence of the documents for the AP task.

3 Social media datasets

For the majority of the experiments we used the corpus gathered by Schler et al. [31][5]. This corpus is a collection of blogs from blogger.com, written in English and collected in August 2004. This corpus is widely used in AP due to its large number of documents (i.e., user profiles) as well as its balanced distribution regarding the number of men and women for each age group. Table 1 shows some numbers about this corpus.

[5] http://u.cs.biu.ac.il/ koppel/BlogCorpus.htm

Table 1. Distribution of the Schler corpus.

Age (age range)	Gender		
	Female	Male	Total
10s (13-17)	4,120	4,120	8,240
20s (23-27)	4,043	4,043	8,086
30s (33-47)	1,497	1,497	2,994
Total	9,660	9,660	19,320

For evaluating the generality of the proposed approach, we used English corpora from different social media domains. For this purpose we considered the corpus from the AP task of PAN-2014[6], referred as PAN-AP-2014, which include data from blogs, reviews, social media and Twitter. As shown in Table 2, all these corpora are balanced regarding gender, but imbalanced regarding age. It is also important to notice that these collections have very different sizes, varying from 147 blog users to 7746 social media profiles.

Table 2. Data distribution of the PAN-AP-2014 corpus.

Corpus	Gender	Age					
		18-24	25-34	35-49	50-64	65 o more	Total
Blogs	Female	3	30	27	11	2	73
	Male	3	30	27	12	2	74
	Total	6	60	54	23	4	147
Twitter	Female	10	44	65	30	4	153
	Male	10	44	65	30	4	153
	Total	20	88	130	60	8	306
Reviews	Female	180	500	500	500	400	2080
	Male	180	500	500	500	400	2080
	Total	360	1000	1000	1000	800	4160
Social Media	Female	775	1049	1123	919	7	3873
	Male	775	1049	1123	919	7	3873
	Total	1550	2098	2246	1838	14	7746

4 Experimental Methodology

This section presents the experimental methodology devised to investigate the relevance of the personal phrases in the AP task. Basically, the central idea of our experiments is to compare the classification performance when using only these phrases vs. the entire documents. Section 4.1 describes the process followed to filter the personal phrases of a document. Then, Section 4.2 details the configuration settings of the classification process used in all the experiments.

[6] http://pan.webis.de/clef14/pan14-web/author-profiling.html

4.1 Filtering Process

We define a personal phrase as a sentence which includes a first person pronoun. We considered the following lists of pronouns: subjective (I, we), objective (me, us), possessive (my, mine, our, ours) and reflexive (myself, ourselves). Second and third person pronouns were not considered because they suggest that the writer is talking about something/someone else without including herself.

The filtering process considers the extraction of all the personal phrases appearing in each document (user history) of a given corpus. As shown in Figure 1, it first splits documents into sentences, and then it selects the sentences which include a first person pronoun. The rest of the sentences, which does not have any personal pronoun, is discarded. In our experiments we refer to these subsets of phrases as the *filtered corpus* and the *complement corpus* respectively. It is important to notice that there could exist documents with no personal phrases, which would lead to empty filtered files. In such situations we decided using the original document instead of the empty filtered file.

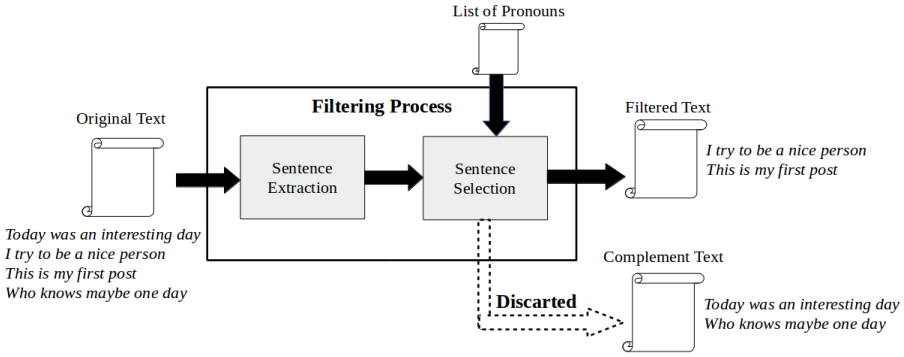

Fig. 1. Filtering Process

4.2 Classification process

For all the experiments we considered a standard classification framework for AP: we used a combination of content and style features, and a Support Vector Machine as learning algorithm [31]. Following we describe the main configuration settings for the classification and evaluation processes.

Features: we used the set of features described in [31, 13]: 1000 content words with the highest information gain, stopwords and punctuation marks, slang words, out-of-dictionary terms like emoticons and POS tags[7].

[7] POS tags were obtained using Stanford tagger: http://nlp.stanford.edu/software/tagger.shtml

Representation: based on all these features, we build a standard BOW representation. The weighting of terms corresponds to their normalized frequency with respect to the total number of terms in the document.

Classifier: To classify the documents, we used the SVM classifier from the LIBLINEAR library [7] without any parameter optimization.

Evaluation: we applied a stratified 10 cross fold validation (10CFV) on each corpus, and used the accuracy as main evaluation measure, which represents the percentage of users that were correctly classified. To assess the statistical differences among the different corpora configurations (original, filtered and complement), we applied a 10CFV paired t test[6, 5].

5 Results and discussion

5.1 Experiment 1: the relevance of personal phrases for AP

The aim of this experiment was to determine the value of the personal phrases for the AP task. Based on the idea that people better expose their interests and writing style when they talk about themselves, this first experiment focused on evaluating the role of the phrases which contain *singular first person pronouns*.

For carrying out this evaluation we used the Schler corpus (refer to Section 3). First, we filtered the personal phrases that contain one of the following pronouns: I, me, mine, my, myself, as well as the string "im", because it is commonly used in social media documents. Table 3 shows some numbers from the resulting corpora. The obtained filtered corpus represents 48.12% of the information of the original collection, and it is smaller than the complement corpus.

Table 3. Data and accuracy results from the first experiment. The filtered corpus is the subset of sentences including singular first person pronouns from the Schler corpus.

	Sentences	Empty files	Age	Gender
Original corpus	9,155,301	0	77.49	80.07
Filtered corpus	4,405,783	69	76.09	79.63
Complement corpus	5,510,302	131	69.98	72.59

To assess the relevance of the personal phrases in AP, we compared the classification accuracy in the age and gender prediction tasks when using the three different corpora. The last two columns of Table 3 show the obtained results. It is worth noting that results obtained using the filtered corpus are significantly better than those corresponding to the complement corpus, even thought there is less information in the former one. This indicates that self-information is indeed more useful for AP than general impersonal information. Furthermore, these results also show that using only the personal phrases it is possible to achieve a very similar performance than using the complete documents. In fact, for the gender prediction there is no statistically significant difference between the results using the filtered and the original corpora. On the one hand, these results

confirm the relevance of the personal phrases for the AP task, and on the other hand, they support our hypothesis that these phrases can be considered as the essence of the documents for this task.

5.2 Experiment 2: the added value of plural personal phrases

The purpose of this experiment was to examine the role of the phrases with *plural first person pronouns* in the AP task. Particularly, it focused on investigating if these phrases, which have inclusive nature and they express information about the user as part of a group, could enrich the representation of users, and consequently could improve their automatic classification.

As in the previous experiment, we used the Schler corpus as reference collection. However, in this case, we considered personal phrases not only containing singular pronouns but also plural first person pronouns. Accordingly, in the filtering process we extracted sentences containing one of the following pronouns: we, us, our, ours, ourselves. Some numbers from the obtained corpora are shown in Table 4. It is worth noting that there are considerably less phrases with plural first person pronouns than with singular first person pronouns, which could be explained by the kind of information shared in blogs. In addition, it can be noticed that their combination only caused an increment of 537,607 phrases (5.9%) over the singular filtered corpus, indicating the frequent co-occurrence of singular and plural first person pronouns in social media posts.

Table 4. Accuracy results using singular and plural personal phrases.

	Sentences	Empty files	Age	Gender
Original corpus	9,155,301	0	77.49	80.07
Singular/plural filtered corpus	4,943,390	33	76.99	79.82
Plural filtered corpus	908,815	1075	67.00	70.35
Singular filtered corpus	4,405,783	69	76.09	79.63

Table 4 shows the accuracy results obtained by the different configurations of the filtered corpus. One first thing to notice is that results corresponding to the use of only singular personal phrases considerably outperformed those obtained by the plural personal phrases. The differences were of 9.1% and 9.3% for age and gender respectively. These differences could be attributed to the difference in the sizes of the corpora, but they also suggest that plural personal phrases change their focus from the user particular interests to the group's concerns.

On the other hand, the test of statistical significance indicated that the observed accuracy differences between the singular/plural filtered corpus and the singular filtered corpus were not statistically significant for both, age and gender, prediction tasks. These results allow us to conclude that plural personal phrases have no special relevance for the AP. Moreover, they also corroborate the outstanding usefulness of the singular personal phrases for this task.

5.3 Experiment 3: content and style information in personal phrases

Previous experiments have shown the important role of personal phrases for the AP task. The purpose of this experiment was to understand the discrimination power of these phrases. Particularly, we wanted to determine the contribution of content and style information from these phrases for the profiling of authors.

For this experiment we divided the features (refer to Section 4.2) into three disjoint sets: *words*, which represent content information, and *function words* and *POS* that represent style information. To assess the relevance of each feature type we compared their classification accuracy when using the singular filtered and complement corpora. Table 5 shows the obtained results.

Table 5. Accuracy results for feature type. The filtered corpus is the subset of sentences including a singular person pronoun from the Schler corpus.

Type of feature	Corpus	Accuracy	
		Age	Gender
Words	Original	76.06	78.12
	Filtered	75.04	78.08
	Complement	68.49	71.19
Function words	Original	68.56	73.05
	Filtered	67.00	70.78
	Complement	61.31	67.56
POS	Original	63.09	68.11
	Filtered	62.87	66.35
	Complement	59.79	65.68

Results from Table 5 confirm conclusions from previous works [31], which have pointed out that content information is more relevant than style information for AP. They also show that the performance difference between the original and filtering corpora is lower in the word space, demonstrating that thematic interests are adequately captured in personal phrases. On the other hand, by comparing the results from the filtered and complement corpora, it is possible to observe an average difference of 6.7% in favor of the filtered corpus when words were used as features, whereas the differences were around 4.4% and 1.9% when using function words and POS features respectively. These results suggest that the value of personal phrases lies mostly in the content aspect rather than in the style information. Hence, we can conclude that style information from authors could be equally well captured from personal and non-personal phrases, nonetheless, topics of interest are better extracted from personal phrases.

5.4 Experiment 4: personal phrases in different social media

The purpose of this experiment was to evaluate the relevance of personal phrases for AP across different social media domains. Mainly, we aimed to corroborate

the generality of our previous findings and check their degree of domain independence. For this experiment we used the PAN-AP-2014 corpus. We built the filtered corpus by selecting the posts that contain singular personal pronouns as detailed in Section 5.1. Table 6 shows some numbers on the obtained corpora.

Table 6. Data from the PAN-AP-2014 corpus. The filtered corpora correspond to the subsets of posts containing a firs person pronoun.

Collection	Posts in original corpus	Posts in filtered corpus	Empty Files
Blogs	22,994	5,565	10
Twitter	318,691	49,540	7
Reviews	52,833	19,248	1,377
Social media	3,207,509	736,615	1,349

Table 7 shows the results across different social media domains. For all the collections we approached two classification problems: age prediction with five classes (18-24, 25-34, 35-49, 50-64, 65 o more), and gender classification with two classes (male and female). The results are very interesting since they present similar accuracy values when using the filtered and the original corpus, although the filtered corpora only represent a small subset (ranging from 15% to 36%) of the original corpora. Particularly, the statistical significance test indicated that results for age prediction were comparable across all considered domain, whereas for the gender classification we found a statistically significant difference for the Twitter and Blog domains. However, it is important to notice that for these two collections we obtained better age prediction results using the filtered corpus than using the original corpus, which causes a comparable overall performance.

Table 7 shows the results across different social media domains. For all the collections we approached two classification problems: age prediction with five classes (18-24, 25-34, 35-49, 50-64, 65 o more), and gender classification with two classes (male and female). The results are very interesting since they present similar accuracy values when using the filtered and the original corpora, although the first only represent a small subset (ranging from 15% to 36%) of the original corpora. Particularly, the statistical significance test indicated that results for age prediction were comparable across all considered domains. This is a very encouraging result since age prediction in these collections considers five age categories with consecutive values and, therefore, it represents a harder classification problem than that from the Schler corpus. On the other hand, for the gender classification we found a statistically significant difference for the Twitter and Blog domains. However, it is important to notice that for these two collections we obtained better age prediction results using the filtered corpus than using the original corpus, which causes a comparable overall performance. In general, these results support the relevance of the personal phrases as well as their role as the essence of the documents for the AP task.

Table 7. Accuracy results at PAN 2014 collections.

Collection	Corpus Conf.	Accuracy		% kept in filtered corpus
		Age	Gender	
Blogs	Original	36.56	68.42	24.20% (from 22,944 posts)
	Filtered	43.92	62.14	
Twitter	Original	35.33	71.33	15.54% (from 318,691 posts
	Filtered	37.49	59.55	
Reviews	Original	30.84	67.24	36.43% (from 52,833 posts)
	Filtered	29.21	65.21	
Social Media	Original	34.84	53.64	22.97% (from 3,207,509 posts)
	Filtered	33.99	52.68	

6 Conclusions and future work

Inspired on the idea that people best reflect their personal characteristics and writing style when they talk about themselves, in this work we investigated the relevance of personal phrases for the author profiling task. The experiments carried out clearly indicated that personal phrases have a huge value for predicting age and gender of social media users, since considering only this kind of phrases we obtained reductions of up to 60% of the information in the user documents and a comparable performance than using all available data. Hence, personal phrases can be considered as the *essence of documents* for the AP task.

Throughout the paper, we answered the research questions outlined in the introduction, finding that: 1) not all the information from a document is equally relevant for this task, personal phrases are more discriminating than non-personal phrases; 2) although plural personal phrases have inclusive nature, they have not a special relevance for the AP task, and their information is not complementary to that from the singular personal phrases; 3) personal phrases better capture content information (user interests), whereas style information can be equally extracted from both personal and non-personal phrases; 4) the relevance of personal phrases is a general characteristic that was observed in different social media domains.

The achieved results motivate us to evaluate the proposed approach in other profiling tasks such as personality identification, as well as to evaluate its appropriateness in other languages, particularly in those where the use of subjective pronouns is uncommon (pronoun-dropping languages). On the other hand, the obtained conclusions encourage us to explore new ideas for taking advantage of the information from personal phrases in the AP task. In particular, we consider that our findings could be applied to design new strategies for constructing corpora, a task highly expensive in terms of effort and time. They also could help the design of novel feature selection methods, as well as new terms and instances weighting schemes.

Acknowledgments. This work was supported under CONACYT project no. 247870 and scholarship 243957.

References

1. Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J.W.: Lexical Predictors of Personality Type. In: Joint Annual Meeting of the Interface and the Classification Society of North America. St. Louis MI (2005)
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically Profiling the Author of an Anonymous Text. Communications of the ACM, 52(2), 119-123 (2009)
3. Cappellato, L., Ferro, N., Jones, G., San-Juan, E. (eds.): CLEF 2015 Labs and Workshops, Notebook Papers, Toulouse, France, September (2015)
4. Chung, C.K., Pennebaker, J.W.: The Psychological Functions of Function Words. In: Fiedler, K. (ed.) Social Communication: Frontiers of Social Psychology, pp. 343-359 (2007)
5. Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research, 7, 1-30 (2006)
6. Dietterich, T.: Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 10, 1895-1923 (1998)
7. Fan, R.E., Chang, K.W., Hsieh, C. J., Wang, X.R., Lin, C.J.: LIBLINEAR: A Library for Large Linear Classification. The Journal of Machine Learning Research, 9, 1871-1874 (2008)
8. Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Valencia, Spain, September, (2013)
9. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric Analysis of Bloggers' Age and Gender. In: Third International ICWSM Conference, pp. 214-217 (2009)
10. Kacewicz, E., Pennebaker, J.W., Davis, M., Moongee, J., Graesser, A.C.: Pronoun Use Reflects Standings in Social Hierarchies. Journal of Language and Social Psychology, 33, 125-143 (2013)
11. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing, 17(4), 401-412 (2002)
12. López-Monroy, A.P., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Villatoro-Tello, E.: INAOE's participation at PAN'13—Notebook for PAN at CLEF 2013: Author Profiling task. In: Forner et al. [8]
13. López-Monroy, A.P., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Stamatatos, E.: Discriminative Subprofile-Specific Representations for Author Profiling in Social Media. Knowledge-Based Systems. 89, 134-147 (2015)
14. Maharjan, S., Solorio, T.: Using Wide Range of Features for Author profiling—Notebook for PAN at CLEF 2015. In: Cappellato et al. [3]
15. Meina, M., Brodzínska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., Wilk, M.: Ensemble-based Classification for Author Profiling Using Various Features—Notebook for PAN at CLEF 2013. In: Forner et al. [8]
16. Mihalcea, R., Hassan, S.: Using the Essence of Texts to Improve Document Classification. RANLP 2005. Borovetz, Bulgaria (2005)
17. Mukherjee, A., Liu, B.: Improving Gender Classification of Blog Authors. In: Conference on Empirical Methods in Natural Language Processing (EMNLP '10). Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 207-217 (2010)
18. Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker, J.W.: Gender differences in language use: an analysis of 14,000 text samples. Discourse Processes, 45, 211-236 (2008)
19. Newman, M., Pennebaker, J., Berry, D., Richards, J.: Lying Words: Predicting Deception from Linguistic Styles. Personality and Social Psychology Bulletin, 29, 665-675 (2003)

20. Nguyen, D., Smith, N.A. Rosé, C.P.: Author Age Prediction from Text using Linear Regression. In: 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities. Association for Computational Linguistics. pp. 115-123 (2011)
21. Pennachiotti, M., Popescu, A.M.: Democrats, Republicans and Starbucks Aficionados: User Classification in Twitter. In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA. pp. 430-438 (2011)
22. Pennebaker, J.: The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury USA (2011)
23. Pennebaker, J., Stone, L.: Words of Wisdom: Language Use Over the Life Span. Journal of Personality and Social Psychology, 85, 291-301 (2003)
24. Rangel, F., Celli, F., Rosso P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Cappellato et al. [3]
25. Rangel, F., Rosso, P.: Use of Language and Author Profiling: Identification of Gender and Age. In: Workshop on Natural Language Processing and Cognitive Science, NLPCS-2013. Marseille, France (2013)
26. Rangel, F., Rosso, P. On the Multilingual and Genre Robustness of EmoGraphs for Author Profiling in Social Media. In: 6th Int. Conf. of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction, CLEF 2015, Springer-Verlag, LNCS(9283), pp. 274-280 (2015)
27. Rangel, F., Rosso, P. On the Impact of Emotions on Author Profiling. In: Information Processing & Management, 52(1), 73-92 (2016)
28. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G. Overview of the Author Profiling Task at PAN 2013. In: Forner et al. [8]
29. Rao D., Yarowsky D., Shreevats A., Gupta M.: Classifying Latent User Attributes in Twitter. In: Proceedings of SMUC-10. pp. 710-718 (2010)
30. Rude, S., Gortner, E.M., Pennebaker, J.W.: Language Use of Depressed and Depression-Vulnerable College Students. Cognition and Emotion, 18, 1121-1133 (2004)
31. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of Age and Gender on Blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. pp. 199-205. AAAI (2006)
32. Schwartz, H.A., Eichstaedt, J.C., Dziurzynski, L., Kern, M.L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H.: Toward Personality Insights from Language Exploration in Social Media. In: AAAI Spring Symposium: Analyzing Microtext. AAAI (2013)
33. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1-47 (2002)
34. Sidorov, G., Miranda Jiménez, S., Viveros Jiménez, F., Gelbukh, A., Castro Sánchez, N., Velásquez, F., Díaz Rangel, I., Suárez Guerra, S., Treviño, A., Gordon, J. Empirical study of opinion mining in spanish tweets. LNAI, 7629-7630 (2012)