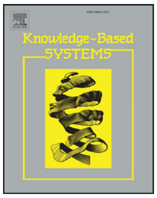
Knowledge-Based Systems 145 (2018) 169–181

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Emphasizing personal information for Author Profiling: New approaches for term selection and weighting

Rosa María Ortega-Mendoza[a,d,*], A. Pastor López-Monroy[b], Anilu Franco-Arcega[a], Manuel Montes-y-Gómez[c]

[a] Universidad Autónoma del Estado de Hidalgo (UAEH), Carr. Pachuca-Tulancingo Km. 4.5, Mineral de la Reforma, Hidalgo, C.P. 42090, Mexico
[b] University of Houston, 4800 Calhoun Road, Houston, Texas, C.P. 77004, USA
[c] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla,Puebla, C.P. 72840, Mexico
[d] Instituto Tecnológico Superior del Oriente del Estado de Hidalgo (ITESA), Carr. Apan-Tepeapulco Km. 3.5, Apan, Hidalgo, C.P. 43900, Mexico

ARTICLE INFO

Article history:
Received 6 June 2017
Revised 19 December 2017
Accepted 17 January 2018
Available online 31 January 2018

Keywords:
Author profiling
Feature selection
Term weighting
Personal information
PEI

ABSTRACT

The Author Profiling (AP) task aims to predict specific profile characteristics of authors by analyzing their written documents. Nowadays, its relevance has been highlighted thanks to several applications in computer forensics, security and marketing. Most previous contributions in AP have been devoted to determine a suitable set of features to model the writing profile of authors. However, in social media this task is challenging due to the informal communication. In this regard, we present a novel approach, which considers that terms located in phrases exposing personal information have a special value for discriminating the author's profile. The aim of this research work is to emphasize the value of such *personal phrases* by means of two new proposals: a feature selection method and term weighting scheme, both based on a novel measure called Personal Expression Intensity (PEI) which scores the quantity of personal information revealed by a term. For evaluating the latter ideas, we show experimental results in age and gender prediction of media users on six different collections. Average improvements of 7.34% and 5.76% for age and gender classification were obtained when comparing to the best result from state-of-the-art, indicating that personal phrases play a key role for the AP task by means of selecting and weighting terms.

1. Introduction

The Author Profiling (AP) task consists in analyzing texts to predict general or demographic attributes of authors such as: gender, age, personality, native language, political orientation, among others. Recently AP has gained a lot of interest because of its applications in areas such as marketing, where companies leverage online reviews to improve targeted advertising, and forensics, where the linguistic profile of authors could be used as valuable additional evidence.

Broadly speaking, AP has been approached as a single-label classification problem [1]. Consequently, most work has been devoted to determine a suitable set of features for modeling the writing profile of authors. In the case of social media documents, AP faces extra challenges derived from the informal communication [2,3]; for example, texts tend to contain grammatical errors, abbreviations, slang words, and even sometimes texts are spurious or automatically generated by bots. All these particularities of social media texts have hindered the direct use of several advanced Natural Language Processing (NLP) tools such as POS-taggers, syntactic and semantic parsers, and have consolidated the use of lexical features as a standard representation approach, which has been broadly used despite of its simplicity [4–6]. Recently, more sophisticated approaches have been considered, yielding good results; for example, using character, word or syntactic n-grams [7,8], topic-based representations [9], second order attributes [10] and word embeddings representations [11].

Although lexical features have demonstrated to be useful for the AP task, little attention has been paid to emphasize the features related to personal information[1], even though recent works in social psychology [12–17] have demonstrated that self-

* Corresponding author.
E-mail addresses: or300944@uaeh.edu.mx (R.M. Ortega-Mendoza), alopezmonroy@uh.edu (A.P. López-Monroy), afranco@uaeh.edu.mx (A. Franco-Arcega), mmontesg@inaoep.mx (M. Montes-y-Gómez).

[1] By personal information we meant the interests, preferences, habits, and any other demographic aspect useful to identify an individual or her membership to a group.

https://doi.org/10.1016/j.knosys.2018.01.014
0950-7051/© 2018 Elsevier B.V. All rights reserved.

references reflect important thematic and stylistic preferences about authors. Based on this observation we hypothesize that personal phrases -sentences containing a first person pronoun[2] better reflect the interests, opinions and feelings of authors [18], and that emphasizing the role of their terms could lead to significant improvements in the AP performance. Following this idea, in this paper we propose two methods for term selection and weighting that are especially suited to profiling social media users. These methods go beyond traditional approaches that exclusively consider the frequency of terms by quantifying the personal information revealed by each of them. In summary, the contributions of this paper are threefold: First, a measure named Personal Expression Intensity (*PEI*), which aims to score the personal information revealed by each term by considering their co-occurrences with first-person pronouns. Second, a new feature selection method that takes advantage of the *PEI* to determine the terms that are simultaneously more descriptive (i.e., personal) as well as discriminative. Third, a new term weighting scheme that uses the *PEI* to boost the relevance of terms that are more associated to the interests of the authors.

The ideas of this paper were evaluated in age and gender prediction using six collections from different domains: blogs, twitter, social media and reviews. The experimental results confirmed that the performance on age and gender identification can be improved by emphasizing the value of the terms highly associated to personal phrases. By using the proposed approach we achieved average accuracy improvements of 7.34% and 5.76% for age and gender classification, respectively, over state of the art results.

The remainder of the paper is organized as follows. Section 2 presents the related work. Section 3 describes the importance of personal phrases for AP task. Section 4 introduces the Personal Expression Intensity score (*PEI*). Section 5 describes the two methods that integrate the proposed approach: a new term selection method and a novel term weighting scheme. Section 7 shows the experiments and discusses the results. Finally, Section 8 exposes our main conclusions.

2. Related work

As previously mentioned, the AP task consists in analyzing texts to predict general or demographic attributes of authors such as their gender [19–22], age [20–23], personality [3,24], native language [20], political orientation [25], among others.

Traditionally, the AP task has been tackled from a text classification perspective [1] by means of the Bag-of-Words (BoW) representation. In this scenario, the focus of research has been on the selection of the best textual features for modeling the authors' writing profile. Two kinds of textual features have been playing a key role: i) content-based features (e.g., word n-grams and topic models), and ii) style-based features (e.g., function words, punctuation marks and emoticons). According to the PAN[3] evaluation forums [4–6], most successful works for AP in social media have used combinations of these two kinds of features. Other works in this direction, such as the work in [20] has used content and style features to identify the age, gender, native language and neuroticism level of authors. [26] studied the classification of blogs by gender using POS patterns as features. Other proposals include the use of stylometric characteristics. For example, [2] predicted age and gender of blogs' authors by means of slang words and the length of sentences, whereas [27] and [28] used style and structural features such as the frequency of capital letters, words length, and number of words with flooded characters (e.g., Heeeellooo), number of sentences, paragraphs, special characters, among others. More recently, some works have used different deep learning models and strategies to learn representations for AP, for example, Neural Attention Models [29], Bidirectional Recurrent Neural Networks [30], Subwords embeddings [31], and Convolutional Neural Networks [32]. It is interesting to point out that most of these works have confirmed that content features (e.g., words) perform better than style features (characters, char *n*-grams, etc.) [33]. Notwithstanding the novelty and complexity of the used strategies, none of these works have outperformed the results obtained by traditional approaches using a combination of Bags-of-Terms (e.g., words, *n*-grams, etc.) and Linear Support Vector Machines [33].

In spite of the relevance of these two kinds of features, social media imposes further challenges to current AP methods. The diversity of the information shared through this media as well as its informal nature lead to huge vocabularies with lot of noise. To face this issue, recent works –from psychological and computational perspectives– are considering different approaches for selecting the most informative features. The following subsections describe some of these works.

2.1. Personal information in Author Profiling

The AP task is based on the idea that people with common profile characteristics also share linguistic similarities, in part because of their social or cultural environments. Works from psychology have studied how language is shared by people [34,35], and they have established a relation between language usage and personality traits [36,37] and gender differences [38] among others. These works have motivated several –computational– studies for AP, which have found that the usage of some function words is strongly related to the expression of feelings, opinions, fears and interests [13]. Furthermore, they have found that patterns on the use of personal pronouns are very useful features to distinguish among different groups of people [12]. For example, the frequent use of singular first-person pronouns is related to: young people [35], females [15,20], low social status [16] and depression [17].

Supported on these ideas, in a previous work [18] we studied the role of personal phrases (phrases containing a first-person pronoun) in AP, demonstrating that they are most valuable than the non-personal phrases for this task. In this paper we move a step forward by proposing new techniques for feature (term) selection and weighting based on their occurrences in personal phrases.

2.2. Feature selection and weighting in Author Profiling

There is a number of feature selection methods that have been successfully used in text classification tasks [39,40]. Nevertheless, for the AP task the most used strategy is by far the frequency threshold of words [10,11,19–21,41]. This is not surprising because it is well known that valuable style features commonly have high frequencies and therefore they tend to be removed by many feature selection methods.[4] The information gain has also been used, but most of the time to analyze or interpret the used features [21], or to extract thematic terms in systems with multiple kinds of features [18,42]. Other efforts for feature selection in AP consider the use of χ^2 [43], point-wise mutual information [3], and combinations of traditional wrapper and filter strategies [26].

[2] Namely: I, me, mine, myself, my, as well as im, which is very common in social media.

[3] http://pan.webis.de/.

[4] Although style information is not the cornerstone in the AP task, it plays an important role for detecting some profiles, for example, emoticons are highly useful for age detection.

The story is not very different for feature weighting. Regardless of the contribution of many weighting schemes for text classification (see for example, [44–46]), there are not suitable proposals for the AP task. In fact, as noticed in recent AP forums [11,41–43], traditional term weightings such as the boolean weight, normalized TF and TFIDF are the most commonly used approaches.

In spite of the heterogeneity of the selection and weighting approaches, the vast majority are supported on statistical inferences about term occurrences in the documents, without considering any qualitative characteristic of these occurrences. In contrast, the approaches proposed in this paper are based on the idea that not all information in a document is equally relevant, and accordingly the selection and weighing of terms is done by taking advantage of their occurrences in personal phrases.

3. On the relevance of the personal information for Author Profiling

As described in Section 2, personal pronouns have shown to be very useful features to distinguish among different groups of people: young people [35], female [15,20], low social status [16], and depression [17].

Based on these findings, in a previous work [18] we studied the role of personal phrases (i.e., phrases containing *singular first-person pronouns*) in the prediction of age and gender of social media users. The pronouns considered to define a sentence as personal were: *I, me, mine, myself, my* and *Im*, which is very popular in social media. We did not consider slang expressions such as AFAIK (“as far as i know”), TIL (“Today I Learned”), or IDK (“I don’t know”), since all of them were infrequent and tended to add a lot of noise into the text representations.[5] For example, AFAIK occurred only 13 times in our corpora, the usage of TIL was more related to “until” than to ”as far as I know” as in “breakfast being served till 1030 am...”, and expressions like IDK showed a high co-occurrence with the pronoun I, such as in posts like “idk most people say i have great hair ...”.

For our analysis, we mainly compared the performance of a state-of-the-art method [21] when it uses the complete documents[6] and only the subset of personal phrases. The results were astonishing; similar accuracy results were obtained using either of the two corpora, although the subset of personal phrases represents only a small portion (from 15% to 50% depending on the social media domain) of the original dataset. Our main conclusion was that personal phrases are the *“essence”* of the documents for the AP task.

In this paper we further develop the idea of using the personal phrases as main information for AP. We presume that their relevance is consequence of the kind and clarity of their content: when people talk about themselves they usually expose their interests and concerns. To illustrate this assumption some personal phrases extracted from a blog are shown in Table 1. These phrases describe activities that people tend to do when they wake up in the morning. As noticed each person has his/her own writing style and thematic interests; however, it is possible to find some discriminative patterns. For example, men used to talk more about their breakfast (food) than women, whereas women talk more about personal care, specifically about their hair. On the other hand, young people tend to mention their parents and to write informally (“..”, “&” or “!!”).

In contrast to the personal phrases, the phrases not containing any first-person pronoun, which we refer to as *non-personal phrases*, tend to talk about different events, objects and individuals, and thus, they are more imprecise in capturing relevant profile information about their authors. Table 2 shows some examples of non-personal phrases from the same users of Table 1.

In one of these examples, the 41-year-old female user is talking about a boy (maybe her son, which is a common topic among women of this age range), but she is not clearly revealing part of her profile because she is talking about the interests of the boy rather than her owns. It is also noticed that style features are equally observed in personal and non-personal phrases. For example, the 15 years-old girl uses the expression “..” in both kind of phrases.

Based on these observations we concluded about the usefulness of both kind of phrases for AP, but also about the greatest relevance of the personal ones. Accordingly, we propose a new measure, the *personal expression intensity* (*PEI*), which aims to estimate how much each term is revealing about the profile of an author.

4. Personal expression intensity (PEI)

As mentioned above, the personal information (interest, preferences, habits, among other kind of private data) shared by social media users is highly related to their profile. In order to determine the amount of personal information revealed by each term we propose three new measures, the *personal precision* (ρ), the *personal coverage* (τ), and the *personal expression intensity* (*PEI*), this last being a combination of the former two. These measures resemble ideas from classic IR evaluation measures. Following we formally describe them.

Let d_j be a document and t_i a term occurring in it. S_j, P_j, and N_j represent the set of phrases in d_j and its subsets of personal and non-personal phrases respectively. Accordingly, the function $\#(t_i, X)$ is used to indicate the number of phrases from set X where the term t_i appears.

Personal precision (ρ**)** indicates the concentration of personal information revealed in the context of a term. It is defined as the percentage of personal phrases in the subset of phrases containing the term.

$$\rho(t_i, d_j) = \frac{\#(t_i, P_j)}{\#(t_i, S_j)} \tag{1}$$

Personal coverage (τ**)** indicates the portion of the personal phrases from a document (i.e., the portion of its ”essence”) covered by the term. It can be interpreted as the conditional probability of the occurrence of a term given the set of personal phrases.

$$\tau(t_i, d_j) = \frac{\#(t_i, P_j)}{|P_j|} \tag{2}$$

Although the values of both ρ and τ become greater when the occurrences of t_i in personal phrases increment, their behavior is somehow opposite. For example, a term appearing only once in a document, but in a personal phrase, will have a large ρ, but not necessarily a big coverage (τ) value. On the contrary, a term appearing in the unique personal phrase from a document will have the highest τ value, regardless its occurrences in non-personal phrases. With the aim of having a balance of these two measures we propose the following measure.

Personal expression intensity (*PEI***)** is a combination of ρ and τ as defined by formula (3). It indicates that the more frequent is a term in the personal phrases of a document, and the less frequent in its non-personal phrases, the more revealing is the term about the profile of the document’s author. A term occurring exclusively in all the personal phrases of a document will yield the highest *PEI*

[5] In an early design stage we also analyzed the expressions IC, IDC, IANAL, IANAD, IDC, IKR, ILY, IIRC IWSN, IMHO, and YKWIM, but all of them were quite infrequent in our corpora.

[6] A document corresponds to the concatenation of all texts, posts or tweets, written by a single author.

Table 1
Examples of personal phrases (*taken from the Schler's Corpus* [21]).

Fragment text	Gender	Age
“And then I woke up at 11:00 & took a *shower* & got *dressed*. Then I was gonna fix my *hair* & put on my *makeup* & *mom* said there was no use in goin because it was late anyway.. So I didn't go”	female	15
“I woke up Sunday morning and *cleaned* up the *house*. I have decided not to run away, just yet. Once the *house* was *cleaned* I took a long *bath* and *washed* my *hair* and gave it an intensive *conditioning treatment*.”	female	41
“I woke up, *ate*, and helped *Dad* in the *basement*. Then at *lunchtime* I *ate* again. At two I had this thing at the *library* where they showed you how to make stuff out of duck tape. Most of it I already knew.”	male	13
“Wow what a day! I woke up about 11:30 to a great *break-fast of tacos!! Beef, egg, cheese* and*salsa sauce to be precise, yummmm*”	male	15
“I woke up this morning feeling great. I went to the *kitchen, fried* me a *hamburger patty*, and some *eggs*. There were a few *dishes* that needed to be *washed* so I *washed* them. I came back up stairs, picked up my *room*, and made my *bed*. It is great to be alive and sober.”	male	44

Table 2
Examples of non-personal phrases for the same users in Table 1.

Fragment text	Gender	Age
“*It's* a pretty good *movie*. It's all about *hockey* though.. kinda boring”	female	15
“*The boy is going on an excursion to a science centre*. He is so excited. *He loves anything sciencey and mathematical.* ”	female	41
“*The Guy had oleyed up to grind* and when *he landed on the rail cracked right* in two.”	male	13
“Did my sister yell at him? Smack his ass? Nope. *She said "he didn't do it on purpose!"* and came to his defense, comforting him ”	male	35

value.

$$PEI(t_i, d_j) = 2\frac{\rho(t_i, d_j)\cdot\tau(t_i, d_j)}{\rho(t_i, d_j) + \tau(t_i, d_j)} \quad (3)$$

Analogously to *PEI*, we also formulated the *non-personal expression intensity* (*NEI*), as defined by formula (4). This measure considers the occurrences of terms in the subset of non-personal phrases and it aims to capture the level of association of each term to non-personal information or, in other words, its irrelevance for the AP task.

$$NEI(t_i, d_j) = 2\frac{n\rho(t_i, d_j)\cdot n\tau(t_i, d_j)}{n\rho(t_i, d_j) + n\tau(t_i, d_j)} \quad (4)$$

where $n\rho$ and $n\tau$ represent the non-personal precision and coverage respectively. These two concepts are analogous to ρ and τ, in Formulas 1 and 2 respectively, but having as numerator $\#(t_i, N_j)$, the number of non-personal phrases where the term t_i appears.

5. The proposed approach for AP

The AP task has traditionally tackled as a supervised text classification problem; its goal is to learn a classifier to assign predefined classes (categories of authors) $C = \{c_1, .., c_{|C|}\}$ to a collection of documents $D = \{d_1, ..., d_{|D|}\}$. The construction of this classifier involves the transformation of documents into a suitable representation for machine learning algorithms. This process involves two main stages: *i*) term selection and *ii*) term weighting. The proposed approach focuses on these two stages. Supported on the *PEI* measure it considers a new feature selection method called *discriminative personal purity* (*DPP*), and a new term weighting scheme called exponential rewarding of personal information (*EXPEI*).

5.1. Term selection: discriminative personal purity (DPP)

The goal of feature selection is to detect the subset of most relevant terms for the classification task. The *DPP* is a feature selection approach especially suited for AP in social media. It not only considers the distribution of terms across the categories, as most traditional measures such as information gain do, but it also considers the kind of phrases they appear in. By means of the *PEI* measure it is possible to choose the terms more closely related to the profiles of the social media users.

Formula (5) defines the *DPP* function, which can be used to select a number of more relevant terms. It consists of two components: first, a descriptive factor, defined as the maximum value of the function PP_k (Eq. (6)), that captures the capability of a term to describe personal information of authors belonging to the category (c_k); and second, a discriminative factor, based on the *gini* coefficient (Eq. (7)), which scores the ability of the term to discriminate among the different categories (profiles) of authors. In the following, both discriminative and descriptive factors are described.

$$DPP(t_i) = \max_{k=1}^{|C|}\{PP_k(t_i)\}\cdot gini(t_i) \quad (5)$$

Categorical personal purity as descriptive factor: The personal purity of a term t_i in a category c_k, defined as $PP_k(t_i)$, assesses the personal information captured by the term from the documents belonging to that category (profile). It is computed as the cumulative quotient of the *PEI* and *NEI* of the term over all the documents from profile c_k. In that way, a term having *PEI* values greater than *NEI* values will be rewarded.

$$PP_k(t_i) = \log_2\left(2 + \frac{1}{2}\sum_{d_j\in c_k}\frac{PEI(t_i, d_j) + 1}{NEI(t_i, d_j) + 1}\right) \quad (6)$$

The $PP_k(t_i)$ formula uses additive smoothing to eliminate the divisions by zero, and it applies the logarithm function to reduce the wide scale while preserving the rating of sums, and to handle the imbalance problem (it is very common to have lot of documents from some profiles while very few from others). At the end $PP_k(t_i)$ returns values greater than 1, indicating the level of relevance of the term t_i to *describe* the personal information from profile c_k.

Gini coefficient as discriminative factor: The purpose of the second component of formula (5) is to assess the capability of a term for discriminating documents from different categories of authors. It mainly evaluates the distribution of a term among all categories (profiles). For example, the concentrated presence of a term in only one of the categories indicates its appropriateness to discriminate users from different profiles. On the other hand, the equal distribution of the occurrences of a term in all categories indicates that such term is worthless for AP.

This second factor is implemented by means of the *Gini coefficient*, a well known measure of the concentration or inequality of any distribution. In our case, we analyze the distribution of a term t_i across the set of categories $C = \{c_1, .., c_{|C|}\}$. Although, this coefficient has been calculated as a proportion of area of Lorenz [47], there are alternative ways to estimate it. We used the formula (7) showed by Dixon [48], whose values range from 0 (indicating complete equality) to 1 (denoting complete inequality).

$$gini(t_i) = \frac{1}{\mu \cdot |C|(|C|-1)} \sum_{k=1}^{|C|} (2k - |C| - 1) \frac{\#(c_k, t_i)}{\#(c_k)} \qquad (7)$$

where $\#(c_k)$ and $\#(c_k, t_i)$ indicate the number of documents from profile c_K and the number of documents from this profile containing the term t_i respectively, and therefore, $\frac{\#(c_k,t_i)}{\#(c_k)}$ represents the relative frequency of t_i in c_k. μ denotes the average of the distribution of relative frequencies. It is important to clarify that for the computation of the Gini coefficient, as mentioned in [48], the categories must be sorted according to their relative frequency, from smallest to largest (ascending); that is: $\frac{\#(c_1,t_i)}{\#(c_1)} \leq \frac{\#(c_2,t_i)}{\#(c_2)} \leq \cdots \frac{\#(c_{|C|},t_i)}{\#(c_{|C|})}$.

5.2. Term weighting: exponential rewarding of personal information (EXPEI)

The proposed weighting scheme considers all the terms from the documents, from both personal and non-personal phrases, but it seeks to emphasize the personal information. These ideas were motivated by our previous experiments referred in Section 3, where we noticed that non-personal phrases also contain relevant information for profiling, although they do it in a lesser degree than personal phrases. Particularly, the *EXPEI* scheme proposes an exponential rewarding to the weight of terms occurring in personal phrases. This greatly differs from the selection phase where it is important to consider only the terms with a greater proportion of occurrences in personal phrases than in non-personal phrases, that is, terms with high personal purity or low association to non-personal information.

Summarizing, the design of the *EXPEI* weighting scheme was conducted by the following ideas: *i*) assigning a weight to each term of the document despite its occurrence in the personal phrases, this initial weight must be related to the overall frequency of the term in the document; *ii*) incrementing the weight of terms in proportion to their occurrence in the subset of personal phrases; *iii*) give equal importance to long and short documents. Formula (8) shows the proposed *EXPEI* weighting scheme:

$$w_{ij} = \left(\sqrt{TF(t_i, d_j)}\right)^{1-PEI(t_i,d_j)} \qquad (8)$$

where $TF(t_i, d_j)$ represents the normalized frequency of t_i in d_j, computed as $\frac{\#(t_i,d_j)}{\#(d_j)}$. Here, the square root over the function *TF* is used to increase each value allowing to perceive the changes caused by the exponent (reward); the reward is based on the *PEI* measure causing the weight of a term to increase according to its personal intensity. Fig. 1 illustrates the behavior of the *EXPEI* weighting scheme for different *PEI* (reward) values.

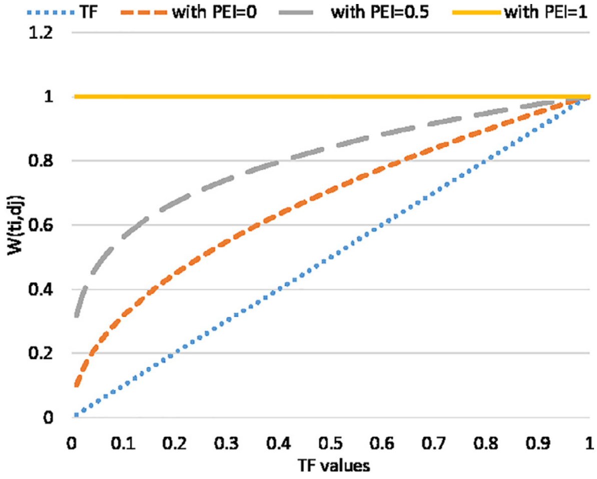


Fig. 1. Plotting of the proposed *EXPEI* term weighting scheme. *EXPEI* is based on the TF values of the terms. For terms having $PEI = 1$, their weights will be equal to one ($EXPEI = 1$, its maximum value), regardless of their frequency. Terms having $0 < PEI < 1$ will be proportionally rewarded; this reward is more important for low frequency terms. Finally, for the terms having $PEI = 0$ the *EXPEI* function smooths their weights (make them a little bit greater than the TF values) allowing low frequent terms to have the opportunity of contributing to description of the document.

6. Datasets

For the evaluation of the proposed method we consider six different social media collections. Each collection includes documents in English and labeled by age and gender categories. The first dataset is a collection of blogs[7][21]. The other five collections were released at the PAN 2014 and 2017 evaluation forums [4,33]. The following sections describe in detail each dataset.

6.1. Schler's corpus

The Schler's dataset [21] is a collection of blogs from blogger.com retrieved in August 2004. It is widely used in AP due to its large number of documents and balanced gender-distribution of documents. It has three labels for the age profile: 10s, 20s and 30s. Regarding the gender profile, it has two labels: male and female. Some statistics about this corpus are shown in Table 3.

Table 3
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

Table 4
Data distribution of the PAN-AP-2014 corpus.

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

6.2. PAN-AP-2014 corpus

This collection contains the training datasets for the AP task used in the PAN 2014 evaluation forum.[8] It consists of four domains, namely, blogs, hotel reviews, social media and tweets. Details about them are shown in Table 4. All these collections are in English and they are balanced regarding to gender, but imbalanced with respect to age. In these collections the number of documents

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

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

Table 5
Distribution of the PAN-AP-2017 English corpus for gender classification.

Gender	Documents
Female	1800
Male	1800
Total	3600

vary from 147 to 7746 social media users. These collections have five labels for age: 18–24, 25–34, 35–49, 50–64 and 65-more.

6.3. PAN-AP-2017 twitter corpus

This collection contains the training dataset for the AP task used in the PAN 2017 evaluation forum.[9] The collection contains Twitter data in the following four languages: English, Spanish, Portuguese and Arabic. The instances for each language are labeled with authors' gender and the specific variation of their native language.[10] For the experiments reported in this paper we only used the English subcorpus for the gender classification task. Some statistics are shown in Table 5.

7. Results and discussion

This section reports the evaluation results of the proposed method in all the collections. Five experiments are presented: *i*) the evaluation of age and gender prediction using the proposed method; *ii*) a detailed analysis of the pertinence of the proposed term selection approach (*DPP*) for AP; *iii*) a detailed analysis of the pertinence of the proposed weighting scheme (*EXPEI*) for AP; and *iv*) the evaluation using different classifiers, and *v*) an analysis of the correlation between the results of AP and the characteristics of the collections.

In all the experiments we used an experimental framework similar to the one used in [10,21]. We considered a combination of content, style and syntactic features. Particularly, we selected the top 1000 terms according to the *DPP* measure. These terms include content words, punctuation marks, slang words and out-of-dictionary terms like emoticons. We also considered the occurrences of stopwords and unigrams of POS tags.[11] By means of these sets of features we built a standard BoW representation and employed the *EXPEI* weighting scheme as well as other traditional term weighting approaches. For the classification phase we used the linear Support Vector Machines (SVM) from the LIBLINEAR library[12] using default parameters [49]. For the evaluation we used a stratified ten fold cross validation (10FCV) approach, the accuracy as evaluation metric in Experiment 1, similar to most work in the state-of-the-art, and the macro balanced F-score (F1) in the rest of the experiments. As main baseline we considered the work by Lopez-Monroy et al. [10] (SSR), which has reported the best results so far for the PAN 2013-2016 collections. We also compared the proposed method and the approaches presented in: [9], which exploits topic-based representations based on Latent Semantic Analysis (LSA) and Linguistic Inquiry and Word Count (LIWC); [50], which presents a method based on information retrieval ideas (IRF); [51], which explores second order attributes (SOA); [52], which uses group-level attributes (GLA) by applying topic analysis methods; [53], which presents an analysis of over 140 million words (MW) of English text drawn from the blogosphere; and [21], where a set of style and content features (SC) is used to find stylistic and content differences between male vs female, and authors of different ages. For the evaluation in the PAN 2017 Twitter collection, we also considered a 70-30% training-test partition, and compared our results against those from the approach presented by our collaborators in [32], which was the top ranked method using word embeddings and Convolutional Neural Networks (CNN) at PAN-2017 [33]. Finally, as advised in [54–56], we evaluated the statistical significance of the obtained results using a 0.05 significance level by means of the Wilcoxon Signed-Ranks (experiments 1 to 3). In each experiment more details about the applied test are given.

7.1. Experiment 1: overall performance evaluation

The purpose of this experiment is to provide a general perspective on the performance of the proposed DPP-EXPEI approach. Indirectly, its goal is to determine the relevance of personal information for the AP task.

Experimental results on the PAN-2014 and Schler collections are shown in Table 6. Additionally, Table 7 shows the results on the Twitter collection from PAN 2017. In general, they show better results for gender than for age identification. This is not surprising given the lower complexity of the gender classification problem, which considers only two categories and has a balanced training set. It is worthy to note that DPP-EXPEI outperformed the baselines methods for the majority of the collections from PAN 2014 as well as for the PAN 2017 gender dataset. Furthermore, it obtained better results than the state-of-the-art approach (SSR) in three out of five collections for the age problem, and in five out of six collections for the gender classification. Particularly, it obtained important gains (for example, a difference of 22% for age classification in the blogs corpus), while losses were small (for example, 4.15% for the age problem in the social media corpus).

We applied the Wilcoxon Signed Rank Test for comparing the accuracy performance of SSR and DPP-EXPEI in the 10 datasets from Table 6. The results indicated that the proposed method is significantly better than SSR at 0.05 significance level. From these results it is possible to conclude that DPP-EXPEI allows focusing in the most relevant terms (personal interests) when having less information (i.e., small collections), whereas, when more textual information is available (bigger collections) DPP-EXPEI has less impact because frequent terms tend to directly expose such personal interests. We also evaluated the difference of these two methods considering their results in the PAN 2017 dataset (Table 7). This extended analysis indicated that SSR and DPP-EXPEI are comparable.

To deeply understand the performance of the proposed approach we conducted an error analysis. Table 8 summarizes the results from this analysis. The *errors* rows represent the total number of samples misclassified from each category, whereas *%nerr* indicates the percentage of errors that were assigned incorrectly to a neighboring category. For example, in category 35–49 of the Blogs collection, the 15 errors were assigned to the 25–34 or 50–64 categories, which corresponds to $nerr = 100\%$. Note that in most of the collections, the majority of the errors correspond to instances wrongly assigned to a neighboring category, indicating, on the one hand, that the age ranges were arbitrarily defined and, on the other hand, that thematic interests and style expressions change with age but in a softly manner.

7.2. Experiment 2: analyzing the DPP selection approach

The purpose of this experiment is to make a further analysis about the individual contribution of the discriminative per-

[9] http://pan.webis.de/clef17/pan17-web/author-profiling.html.

[10] The age dimension was not considered in the 2017 evaluation campaign.

[11] POS tags were obtained using Stanford tagger: https://nlp.stanford.edu/software/stanford-postagger-2013-06-20.zip.

[12] According to [49], this library is very efficient on large sparse data sets such as text classification applications.

Table 6
Accuracy results of DPP-EXPEI for age and gender classification in the PAN-2014 and Schler collections.

		PAN 2014 data sets				Schler corpus
	Approach	Reviews	Twitter	Blogs	Social Media	
Age	DPP-EXPEI	**44.83**	**61.44**	**75.34**	33.91	75.9
	SSR [10]	36.9	49.01	53.06	**38.06**	**77.68**
	LSA [9]	34	39	48	36	–
	LIWC [9]	29	47	42	34	–
	IRF [50]	37.62	52.61	45.58	42.51	–
	SOA [51]	33.92	47.97	48.07	37	–
	GLA [52]	–	–	–	–	72.83
	MW [53]	–	–	–	–	77.4
	SC [21]	–	–	–	–	76.01
Gender	DPP-EXPEI	**76.42**	**81.5**	**84.25**	**58.57**	79.43
	SSR [10]	69.27	71.69	80.95	55.39	**82.01**
	LSA [9]	65	66	70	52	–
	LIWC [9]	62	71	60	50	–
	IRF [50]	71.03	78.76	82.99	57.04	–
	SOA [51]	68.05	71.92	77.96	55.36	–
	GLA [52]	–	–	–	–	75.04
	MW [53]	–	–	–	–	80.5
	SC [21]	–	–	–	–	80.01

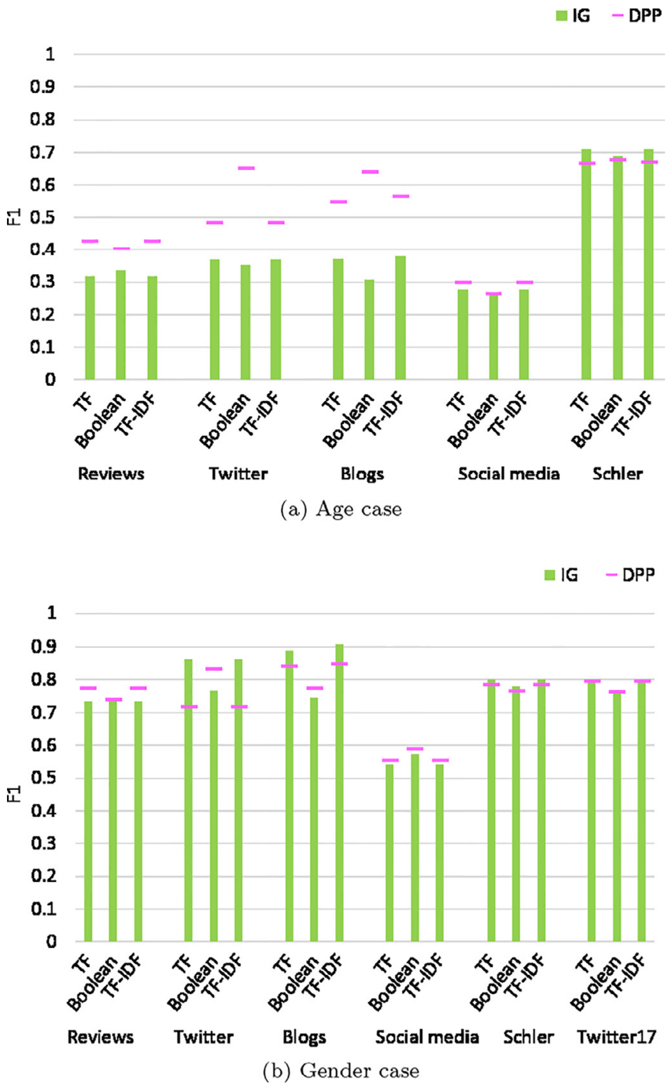
Table 7
Results in the PAN 2017 dataset: gender classification in Twitter.

Approach	Accuracy
10FCV Evaluation Results	
DPP-EXPEI	77.97
SSR [10]	77.65
LSA [9]	76.31
LIWC [9]	69.14
SOA [51]	75.21
70% train, 30% test	
DPP-EXPEI	78.88
CNN-Words [32]	78.05

Table 8
Percentage of misclassified samples assigned to a neighboring category (%nerr). Neighboring categories are the previous or next in the range of age. For example, %nerr 52.60% in 25–34 means that 52.6% of the misclassification are in the 18–24 or 35–49 categories.

Data sets	categories	18–24	25–34	35–49	50–64	65+
Reviews	errors	265	481	589	571	389
	%nerr	47.92	52.60	68.93	64.80	39.85
Twitter	errors	14	30	36	36	2
	%nerr	71.43	86.67	94.44	77.78	0
Blogs	errors	4	6	15	7	4
	%nerr	75.0	66.67	100.0	28.57	0
Social	errors	1017	1389	1398	1302	13
Media	% nerr	39.52	68.9	82.19	50.77	7.69
Data set	categories	10s	20s	30s	–	–
Schler	errors	1119	1903	1634	–	–
	% nerr	85.52	100	88	–	–

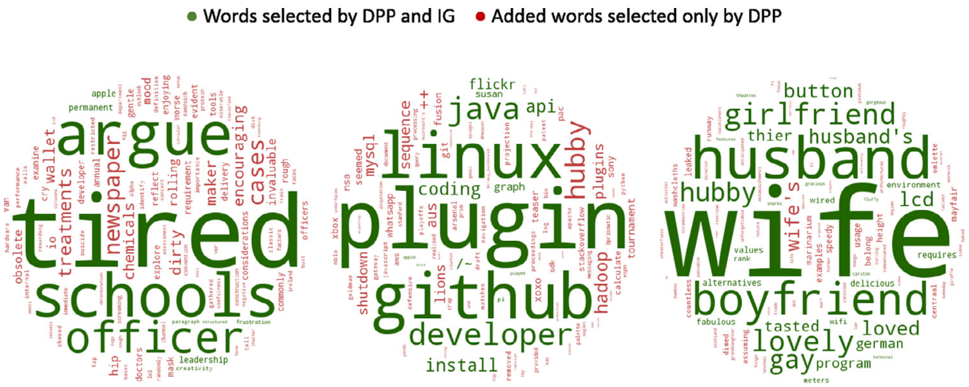
sonal purity (*DPP*) approach to the proposed method. Accordingly, we compared the performance obtained by using two term selection approaches: *DPP* and the traditional information gain (*IG*). For both cases we considered three different term weighting schemes, namely: TF (normalized frequency), boolean and TF-IDF. The results of this comparison are showed in Fig. 2, where columns indicate the f-measure values obtained when *IG* was used for term selection and dashes represent the f-measure values obtained with *DPP*. These results show important improvements for all term weighting schemes when using *DPP*, especially for the age case, and similar performances for the gender classification problem. This sug-



(a) Age case

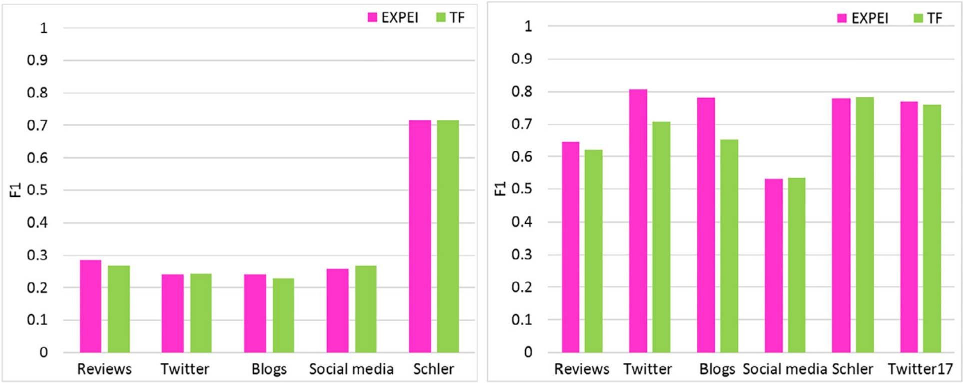
(b) Gender case

Fig. 2. Comparison of the classification performance when discriminative personal purity and information gain are used for term selection.



(a) Blogs: age case (b) Twitter: gender case (c) Reviews: gender case

Fig. 3. Top-100 terms selected by *DPP* for three collections from PAN 2014. The font size is related to the relevance according to *DPP*. The horizontal orientation and green color represent words selected by both *IG* and *DPP*. The red color and vertical orientation indicate the words disregarded by *IG* but relevant for *DPP*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Age case (b) Gender case

Fig. 4. Relevance of the proposed EXPEI weighting scheme. F1 results using the 10,000 top frequent terms with *EXPEI* and *TF* term weights.

gests that personal terms are highly relevant for describing age profiles which are, to some extent, *homogeneous* regardless of gender of the users. Similarly, the low performance for gender classification is not surprising since it is harder to find suitable textual features that applies to very *heterogeneous* profiles (for example, terms that characterize females from five different age ranges), especially when there is few data to extract such evidence (e.g., twitter and blogs datasets are quite small).

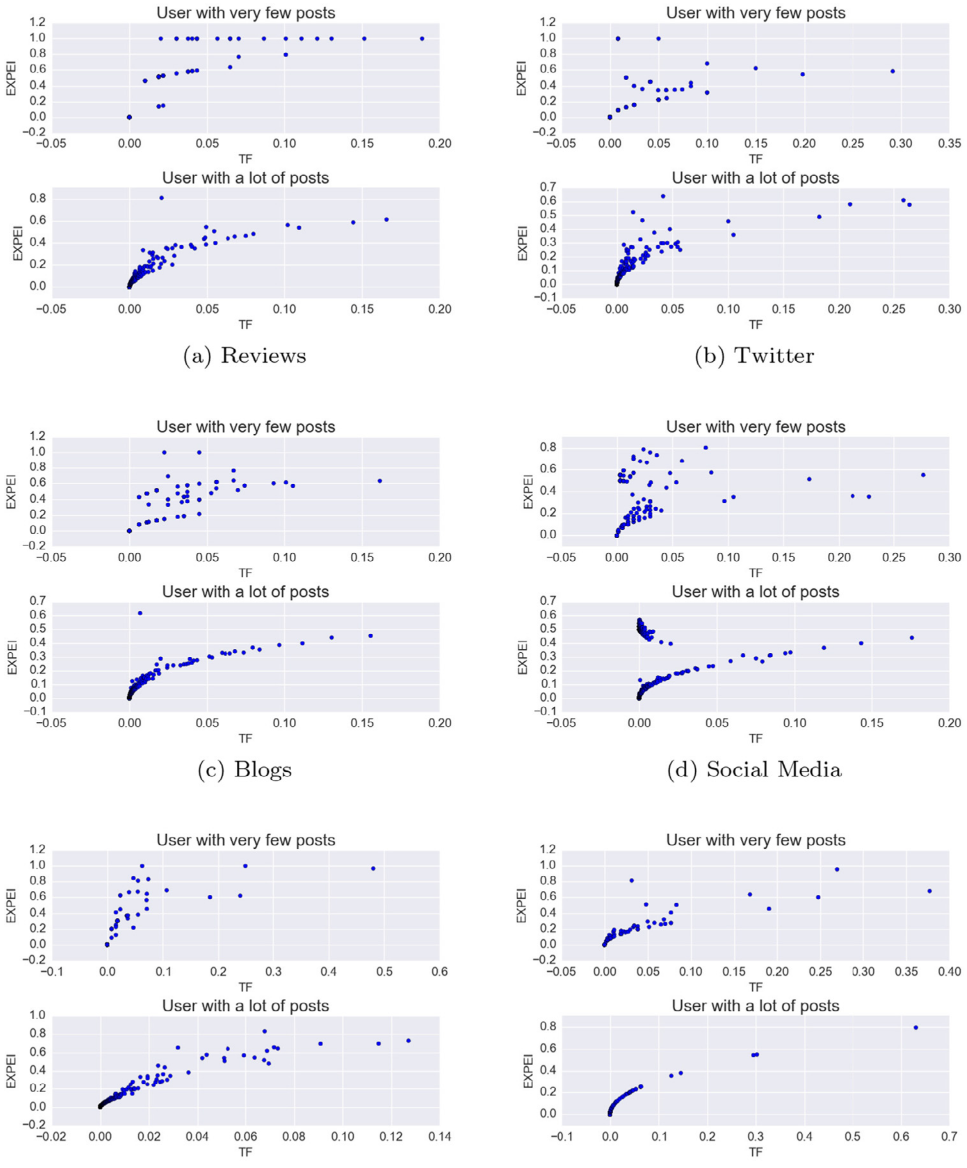
In general, results from Fig. 2 indicate that *DPP* is better than *IG*. According to the Wilcoxon Signed Rank, applied over the three schemes on the 5 collections for age and 6 for gender, *DPP* is significantly better than *IG* with a significance level of 0.05 for the age case, and comparable to *IG* for the gender case. In order to deepen the analysis of *DPP*, Fig. 3 shows a word-cloud of the top 100 selected words by *DPP* corresponding to the following subproblems: blogs-age, twitter-gender and reviews-gender. In the word-clouds, the size of the font corresponds to the rank position. The green color and horizontal orientation represent words selected by both *DPP* and *IG*, whereas the red color and vertical orientation identify the words selected only by *DPP*. From this figure it is possible to observe that the most relevant terms selected by the proposed approach were also selected by *IG*. However, several terms selected only by *DPP* are not frequent but are considered intuitive well-known terms in the AP literature. For example, to identify old people *newspaper, doctors* and *treatments* are valuable terms. For gender identification the words *mysql, hadoop* or *plugins* are closely related to the technology-topic, which has been associated with men. Finally, words like *xoxo, aws, hubby* have shown to be useful for women identification. It is important to note that *DPP* ranks every term in the vocabulary and, therefore, it allows to enrich the selection by including many terms related to personal expressions that usually have zero information gain.

7.3. Experiment 3: analyzing the EXPEI term weighting scheme

It is worth noting that, as shown in Tables 6 and 7, term weighting is crucial to boost the performance of the *DPP* selection approach. For this reason, this section studies the individual contribution of different term weighting schemes. It is mainly focused on analyzing the contribution of the *EXPEI* weighting scheme to the AP task. Particularly, in this experiment we did not use any feature selection technique and considered a representation formed by the top 10,000 most frequent terms in the vocabulary.

Fig. 4 shows the results of the evaluation of the *EXPEI* and *TF* (normalized frequency) weighting schemes. The Wilcoxon Signed Rank Test, applied over the macro F1 results obtained at each fold of both classification problems, indicates that the results are comparable for the age case, whereas for the gender case *EXPEI* significantly outperformed *TF* (with $p < 0.05$), although the more notorious differences are for the gender case and for the smaller collections. These results suggest that males and females tend to use similar vocabularies, and their difference mainly relies on the specific terms used to describe their interests and concerns, which are better captured by *EXPEI* than *TF*.

We also evaluated the influence of the documents' length to the results. We mainly analyze the correlation of weights from the top 10K frequent terms obtained by EXPEI and TF for two kind of users: *i*) users having very few posts and *ii*) users having a lot of



(a) Reviews

(b) Twitter

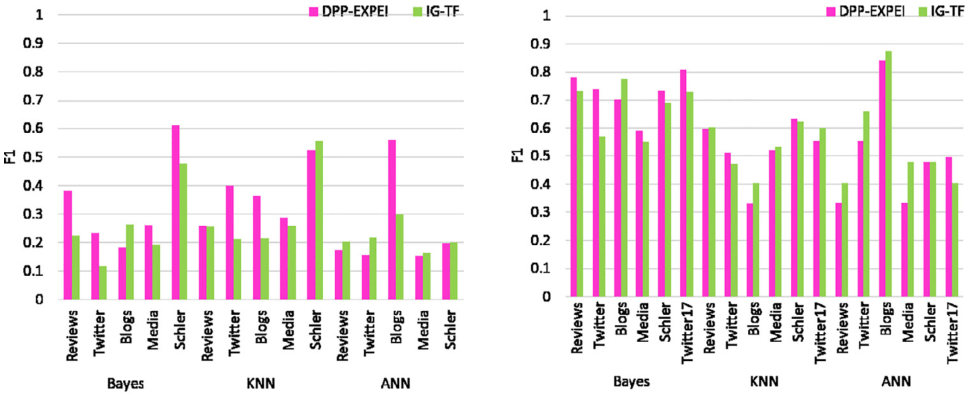
(c) Blogs

(d) Social Media

(e) Schler

(f) Twitter17

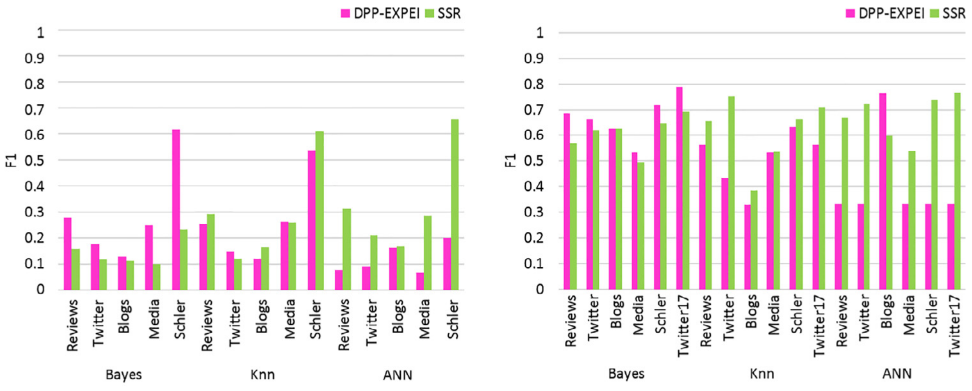
Fig. 5. The correlation of term weights for user profiles having very few and many posts. The graphs show TF and EXPEI correlations between the vectorial representations of the documents. From these results we observe that both representations are less correlated when there are fewer posts, and more correlated in the other case.



(a) Age case

(b) Gender case

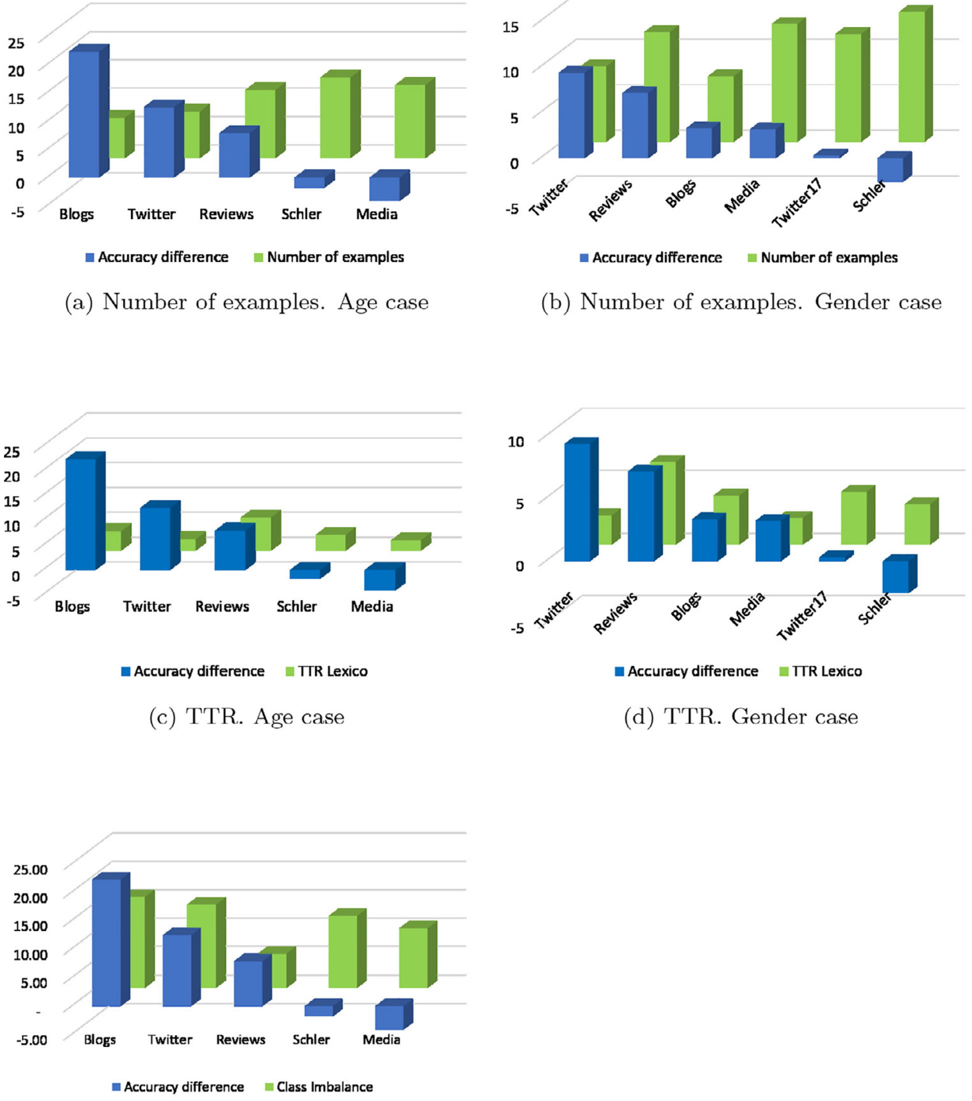
Fig. 6. Comparison of performance using several classification algorithms. A traditional IG-TF is the baseline method.



(a) Age case

(b) Gender case

Fig. 7. Comparison of performance using several classification algorithms. SSR is the baseline method [10].



(a) Number of examples. Age case

(b) Number of examples. Gender case

(c) TTR. Age case

(d) TTR. Gender case

(e) Class imbalance. Age case

Fig. 8. Correlations between some collection properties and the performance of the proposed approach.

posts. Fig. 5 shows the results of this analysis. Note that both representations are less correlated for users having few posts, indicating that our proposal is able to detect infrequent words as relevant features for author profiling. On the other hand, having lots of posts, the correlation between the two approaches tends to increase, suggesting that for long profiles the solely word frequency information is relevant for detecting the most useful features.

7.4. Experiment 4: robustness to different classifiers

The previous experiments have shown improvements when term selection (*DPP*) and weighting (*EXPEI*) are used in conjunction with a Support Vector Machine (SVM). This performance was expected since SVM's consider feature weighting aspects within the learning algorithm. In spite of the effectiveness of SVM, we are interested in studying the robustness of the proposed representation when other classifiers are used. Particularly, we aimed to observe if the proposed method is able to obtain comparable results to other representations, whatever the used classifier.

Figs. 6 and 7 compare the F1 values obtained using the following classifiers with default parameters in Weka 3.8: Multinomial Naïve Bayes (Bayes), 1-Nearest Neighbor (KNN), and an Artificial Neural Network, specifically a perceptron multilayer (ANN). Fig. 6 compares DPP-EXPEI versus a traditional IG-TF approach, it shows that DPP-EXPEI considerably outperformed TF-IDF with all classifiers except ANN. On the other hand, Fig. 7 compares DPP-EXPEI versus SSR using 80% for training and 20% for test. Results suggest that DPP-EXPEI has better results than SSR when using Naive Bayes, comparable performance when using KNN, and tends to have worse results with ANN.[13] In general, the results indicate that: *i*) the proposed method is better or comparable than the baselines using different classifiers; *ii*) the selection of the classifier is important to better exploit the proposed method; and *iii*) the best combination is $DPP - EXPEI$ with a SVM classifier, however, acceptable results can be obtained with different algorithms.

7.5. Experiment 5: on the role of the collection characteristics

The purpose of this experiment is to analyze the role or influence of different characteristics from the collections over the performance of the proposed method. In particular we analyzed the correlation between some specific properties of the collections and the accuracy improvement of the method over the baseline approach (SSR, [10]). For this analysis we show a join plot of the distributions of the obtained improvements and the values of the following properties for the six collections.

- **Number of examples:** it represents the total number of documents in the collection.
- **Type token ratio (TTR):** it is a measure of the vocabulary richness that corresponds to the ratio of different tokens in a given document [57].
- **Class imbalance:** it is calculated as the standard deviation of the differences between the current and the ideal number of documents for each category. The ideal number of documents per category is defined as the ratio of the number of documents in the collection and the number of categories. The higher the value of imbalance class, the more unbalanced the collection is [58].[14]

Fig. 8 shows the obtained plots.[15] They suggest some interesting aspects from the $DPP - EXPEI$ approach. Firstly, it tends to show the highest improvements in collections having the small number of examples. This was especially evident for the age classification problem. Secondly, it also tends to show the highest improvements for collections having the largest imbalance rates.[16] These two characteristics of the proposed approach are very important since it is very common to have imbalanced and small training sets in most AP applications.

8. Conclusions

This paper is supported on the idea that personal phrases (sentences having a first-person pronoun) integrate the essence of texts for the Author Profiling task. That is, personal phrases expose interests, preferences, habits and routines, which help to highlight terms revealing personal traits. In this research work we devised novel strategies to exploit terms occurring in personal phrases by means of term selection and weighting methods. More specifically, three proposals were designed to compose the whole approach. First, the Personal Expression Intensity (*PEI*), a measure aimed to estimate the terms association with personal information. Second, a feature selection method (*DPP*) to select the most valuable terms according to their personal information. And third, the *EXPEI* term weighting scheme, which distinguishes the values of terms occurring in personal and non-personal phrases. The experimental evaluation showed strong evidence about the usefulness of proposed approach. Particularly, the proposed term selection and weighting methods helped to improve the performance in most of the experiments. It is worth to mention that, for age profiling, term selection resulted to be very important, whereas for gender profiling, the term weighting method was more relevant. We consider this performance is because for gender profiling the term weights that reflect term usage are very important, whereas in age profiling the only presence/absence of specific topics are enough to determine the appropriated profile. In spite of the good results obtained by the selecting and weighting methods, their combination outperformed the state-of-the-art in most of the collections showing average accuracy improvements of 7.34% and 5.76% for age and gender cases respectively. The results motivated the evaluation of the proposed approach in other profiling tasks in social media, such as personality identification (e.g., extroverted vs introverted), or other languages, specially in those where the use of subjective pronouns is uncommon (pronoun-dropping languages). Also, we are interested in testing others POS taggers specially suited for Social Media as well as in combining the proposed approach with some deep learning architectures. Finally, we are also interested in exploiting the concept of emphasizing the personal information for scoring terms to weight and rank instances for the AP task.

Acknowledgments

This work was supported by CONACYT under scholarship 243957, and project grants PDCPN2014-01-247870 and FC-2016-2410. The work was also supported by Red Temática en Tecnologías del Lenguaje (projects 260178 and 271622).

[13] This behavior is probably because the proposed representation is of high dimensionality, which makes necessary to perform more parameter tuning.

[14] For our purposes here, we divided the values of the imbalance class by the number of documents in the collection.

[15] The values of the properties were scaled before plotting the distributions using a logarithmic scale factor for the number of examples and a linear factor for the other two properties.

[16] We only analyzed the class imbalance for the age classification problem; the used collections are balanced with respect to the gender categories.

References

[1] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. 34 (1) (2002) 1–47.

[2] S. Goswami, S. Sarkar, M. Rustagi, Stylometric analysis of bloggers' age and gender., in: E. Adar, M. Hurst, T. Finin, N.S. Glance, N. Nicolov, B.L. Tseng (Eds.), ICWSM, The AAAI Press, 2009, pp. 214–217.

[3] H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M.E.P. Seligman, L.H. Ungar, Personality, gender, and age in the language of social media: the open-vocabulary approach, PloS One 8 (9) (2013). E73791

[4] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, W. Daelemans, Overview of the 2nd author profiling task at pan 2014, in: L. Cappellato, N. Ferro, M. Halvey, W. Kraaij (Eds.), CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, Vol. 1180 of CEUR Workshop Proceedings. 15-18 September, CEUR-WS.org, Sheffield, UK, 2014, pp. 898–927.

[5] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, W. Daelemans, Overview of the 3rd author profiling task at PAN 2015, in: L. Cappellato, N. Ferro, G. Jones, E. San Juan (Eds.), CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, CEUR-WS.org, Toulouse, France, 2015.

[6] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, B. Stein, Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations, in: Working Notes Papers of the CLEF 2016 Evaluation Labs, Vol. 1609 of CEUR Workshop Proceedings, CLEF and CEUR-WS.org, 2016.

[7] J.P. Posadas-Durán, H. Gómez-Adorno, I. Markov, G. Sidorov, I.Z. Batyrshin, A.F. Gelbukh, O. Pichardo-Lagunas, Syntactic n-grams as features for the author profiling task: notebook for PAN at CLEF 2015, in: L. Cappellato, N. Ferro, G. Jones, E.S. Juan (Eds.), CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, 2015. Toulouse, France

[8] A. Basile, G. Dwyer, M. Medvedeva, J. Rawee, H. Haagsma, M. Nissim, N-gram: new groningen author-profiling model, 2017, CoRR abs/1707.03764.

[9] M.A. Álvarez-Carmona, A.P. López-Monroy, M.M.-y. Gómez, L.V. nor Pineda, I. Meza, Evaluating topic-based representations for author profiling in social media, in: M.M.y. Gómez, H.J. Escalante, A. Segura, J.d. D. Murillo (Eds.), Advances in Artificial Intelligence - IBERAMIA 2016: 15th Ibero-American Conference on AI, San José, Costa Rica, November 23-25, Springer International Publishing, Cham, 2016, pp. 151–162.

[10] A.P. López-Monroy, M. Montes-y Gómez, H.J. Escalante, L. Villaseñor Pineda, E. Stamatatos, Discriminative subprofile-specific representations for author profiling in social media, Knowl.-Based Syst. 89 (2015) 134–147.

[11] R.K. Bayot, T. Gonçalves, Author profiling using svms and word embedding averages, in: K. Balog, L. Cappellato, N. Ferro, C. Macdonald (Eds.), CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, CEUR-WS.org, Évora, Portugal, 2016.

[12] J. Pennebaker, The Secret Life of Pronouns: What Our Words Say About Us, Bloomsbury, USA, 2011.

[13] C. Chung, J.W. Pennebaker, The psychological function of function words, in: Social Communication: Frontiers of Social Psychology, Psychology Press, 2007, pp. 343–359.

[14] M.L. Newman, J.W. Pennebaker, D.S. Berry, J.M. Richards, Lying words: Predicting deception from linguistic styles, Personality Social Psychol. Bull. 29 (5) (2003) 665–675.

[15] M.L. Newman, C.J. Groom, L.D. Handelman, J.W. Pennebaker, Gender differences in language use: an analysis of 14,000 text samples, Discourse Process. 45 (3) (2008) 211–236.

[16] E. Kacewicz, J.W. Pennebaker, M. Davis, M. Jeon, A.C. Graesser, Pronoun use reflects standings in social hierarchies, J. Lang. Social Psychol. (2013). 0261927X13502654

[17] S. Rude, E.-M. Gortner, J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cognit. Emotion 18 (8) (2004) 1121–1133.

[18] R.M. Ortega-Mendoza, A. Franco-Arcega, A.P. López-Monroy, M.M.-y. Gómez, I, me, mine: the role of personal phrases in author profiling, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings, Springer International Publishing, Cham, 2016.

[19] M. Koppel, S. Argamon, A.R. Shimoni, Automatically categorizing written texts by author gender, Literary Linguist. Comput. 17 (4) (2002) 401–412.

[20] S. Argamon, M. Koppel, J.W. Pennebaker, J. Schler, Automatically profiling the author of an anonymous text, Commun. ACM 52 (2) (2009) 119–123.

[21] J. Schler, M. Koppel, S. Argamon, J. Pennebaker, Effects of age and gender on blogging, in: Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 2006, pp. 199–205.

[22] A.P. López-Monroy, M. Montes-y Gómez, H.J. Escalante, L. Villaseñor Pineda, E. Villatoro-Tello, INAOE's participation at PAN'13: Author profiling task—notebook for PAN at CLEF 2013, in: P. Forner, R. Navigli, D. Tufis (Eds.), CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, 2013. Valencia, Spain

[23] D. Nguyen, N.A. Smith, C.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, Association for Computational Linguistics, 2011, pp. 115–123. Stroudsburg, PA, USA

[24] S. Argamon, S. Dhawle, M. Koppel, J.W. Pennebaker, Lexical predictors of personality type, in: Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America, 2005.

[25] M. Pennacchiotti, A.-M. Popescu, Democrats, republicans and starbucks afficionados: user classification in twitter, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, ACM, New York, NY, USA, 2011, pp. 430–438.

[26] A. Mukherjee, B. Liu, Improving gender classification of blog authors, in: Proceedings of the 2010 conference on Empirical Methods in natural Language Processing, Association for Computational Linguistics, 2010, pp. 207–217.

[27] F. Rangel, P. Rosso, Use of language and author profiling: identification of gender and age, in: In 10th International workshop on natural language processing and cognitive sciences NLPCS 2013 CIRM, October13-17, Marseille, France, 2013, pp. 177–186.

[28] M. Meina, K. Brodzinska, B. Celmer, M. Czoków, M. Patera, J. Pezacki, M. Wilk, Ensemble-based classification for author profiling using various features, in: P. Forner, R. Navigli, D. Tufis, N. Ferro (Eds.), Working Notes for CLEF 2013 Conference,Valencia, Spain, September 23-26 ,2013., CEUR Workshop Proceedings, CEUR-WS.org, 2013.

[29] Y. Miura, T. Taniguchi, M. Taniguchi, T. Ohkuma, Author profiling with word+character neural attention network—notebook for PAN at CLEF 2017, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, CEUR-WS.org, Dublin, Ireland, 2017.

[30] D. Kodiyan, F. Hardegger, S. Neuhaus, M. Cieliebak, Author profiling with bidirectional RNNs using attention with GRUs—notebook for PAN at CLEF 2017, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, CEUR-WS.org, Dublin, Ireland, 2017.

[31] M. Franco-Salvador, N. Plotnikova, N. Pawar, Y. Benajiba, Subword-based deep averaging networks for author profiling in social media—notebook for PAN at CLEF 2017, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, CEUR-WS.org, Dublin, Ireland, 2017.

[32] S. Sierra, M. Montes-Y-Gómez, T. Solorio, F. González, Convolutional neural networks for author profiling in PAN 2017—notebook for PAN at CLEF 2017, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, CEUR-WS.org, Dublin, Ireland, 2017.

[33] F. Rangel, P. Rosso, M. Potthast, B. Stein, Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in twitter, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), Working Notes Papers of the CLEF 2017 Evaluation Labs, Vol. 1866 of CEUR Workshop Proceedings, CLEF and CEUR-WS.org, 2017.

[34] J.W. Pennebaker, M.R. Mehl, K.G. Niederhoffer, Psychological aspects of natural language use: our words, our selves, Ann. Rev. Psychol. 54 (1) (2003) 547–577.

[35] J.W. Pennebaker, L.D. Stone, Words of wisdom: language use over the life span, J. Personality Social Psychol. 85 (2) (2003) 291–301.

[36] L.A. Fast, D.C. Funder, Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior, J. Personality Social Psychol. 94 (2) (2008) 334.

[37] T. Yarkoni, Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers, J. Res. Personality 44 (3) (2010) 363–373.

[38] M.R. Mehl, S. Vazire, N. Ramírez-Esparza, R.B. Slatcher, J.W. Pennebaker, Are women really more talkative than men? Science 317 (5834) (2007) 82.

[39] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 412–420.

[40] G. Forman, An extensive empirical study of feature selection metrics for text classification, J. Mach. Learn. Res. 3 (2003) 1289–1305.

[41] P. Gencheva, M. Boyanov, E. Deneva, P. Nakov, G. Georgiev, Y. Kiprov, I. Koychev, Pancakes team: a composite system of domain-agnostic features for author profiling, in: K. Balog, L. Cappellato, N. Ferro, C. Macdonald (Eds.), CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, CEUR-WS.org., Évora, Portugal.

[42] M. Agrawal, T. Gonçalves, Age and gender identification using stacking for classification, in: K. Balog, L. Cappellato, N. Ferro, C. Macdonald (Eds.), CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, CEUR-WS.org, Évora, Portugal, 2016.

[43] I. Bilan, D. Zhekova, Caps: a cross-genre author profiling system, in: K. Balog, L. Cappellato, N. Ferro, C. Macdonald (Eds.), CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, CEUR-WS.org, Évora, Portugal, 2016.

[44] Y. Liu, H.T. Loh, A. Sun, Imbalanced text classification: a term weighting approach, Expert Syst. Appl. 36 (1) (2009) 690–701.

[45] Y.F. Li, A feature weight algorithm for text classification based on class information, in: Information Technology Applications in Industry, Computer Engineering and Materials Science, Vol. 756 of Advanced Materials Research, Trans Tech Publications, 2013, pp. 3419–3422.

[46] H.J. Escalante, M.A. GarcÃa Limón, A. Morales, M. Graff, M. Montes-y Gómez, E.F. Morales, J.M. Anez Carranza.

[47] M.O. Lorenz, Methods of measuring the concentration of wealth, Publ. Am. Stat. Assoc. 9 (70) (1905) 209–219.

[48] P. Dixon, J. Weiner, T. Mitchell-Olds, R. Woodley, Erratum to bootstrapping the gini coefficient of inequality, Ecology 68 (1988) 1307.

[49] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.

[50] E.R.D. Weren, V.P. Moreira, J. Palazzo M. de Oliveira, Exploring information retrieval features for author profiling—notebook for PAN at CLEF 2014, in: L. Cappellato, N. Ferro, M. Halvey, W. Kraaij (Eds.), CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Vol. 1180 of CEUR Workshop Proceedings, CEUR-WS.org, Sheffield, UK, pp. 1164–1171.

[51] A.P. López-Monroy, M. Montes-y Gómez, H.J. Escalante, L.V. Pineda, Using intra-profile information for author profiling, in: L. Cappellato, N. Ferro, M. Halvey, W. Kraaij (Eds.), CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Vol. 1180 of CEUR Workshop Proceedings, CEUR-WS.org, Sheffield, UK, 2014, pp. 1116–1120.

[52] L.B. Booker, Finding identity group "fingerprints" in documents, in: Computational Forensics: Second International Workshop, IWCF 2008, Washington, DC, USA, August 7-8, Proceedings, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2008.

[53] S. Argamon, M. Koppel, J.W. Pennebaker, J. Schler, Mining the blogosphere: age, gender and the varieties of self-expression, first monday, 2007, http://firstmonday.org/ojs/index.php/fm/article/view/2003. 12, 9.

[54] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[55] S. García, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power, Inf. Sci. 180 (10) (2010) 2044–2064. Special Issue on Intelligent Distributed Information Systems

[56] I. Rodríguez-Fdez, A. Canosa, M. Mucientes, A. Bugarín, STAC: a web platform for the comparison of algorithms using statistical tests, in: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2015. Istanbul (Turkey)

[57] B. Laufer, P. Nation, Vocabulary size and use: Lexical richness in L2 written production, Appl. Ling. 16 (3) (1995) 307–322.

[58] P. Rosso, F.P. Tellez, D. Pinto, J. Cardiff, Defining and evaluating blog characteristics, in: 2013 12th Mexican International Conference on Artificial Intelligence 00, 2009, pp. 97–102.