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/

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.