Social-Media Users can be profiled by their Similarity with other Users

Notebook for PAN at CLEF 2017

A. Pastor López-Monroy[1], Manuel Montes-y-Gómez[2], Hugo Jair Escalante[2], Luis Villaseñor-Pineda[2], and Thamar Solorio[1]

[1]Department of Computer Science,
University of Houston, USA.
alopezmonroy@uh.edu, solorio@cs.uh.edu
[2]Department of Computer Science,
Instituto Nacional de Astrofísica, Óptica y Electrónica, México,
{mmontesg, hugojair, villasen}@inaoep.mx

Abstract. In this paper we describe the system we designed for participating at CLEF-PAN 2017. In this work, we addressed the Author Profiling (AP) task by exploiting the corpus as a knowledge base. The core idea is that profiles can be identified by exposing its relationship with other users. This strategy produces enriched representations, where fine grained user-document relationships are highlighted. For this, we use a non-sparse user-document representation, which captures distributional information of word-usage among user-documents. We compare the proposed approach with the Second Order Attributes, which have been a key component in the winning approaches of the previous PAN-AP editions. We also report experimental results with the traditional Bag-of-Terms and Avg-Word2Vec representations. The experimental evaluation on the PAN17 corpora shows that the proposal outperforms all other methodologies, showing strong evidence of the usefulness of the representation to determine language variety and gender profiles. Furthermore, this representation can be seen as a natural extension to Second Order Attributes, which could be combined in future works in order to expose finer details about user relationships.

1 Introduction

Recently, the Author Profiling (AP) task has gained the interest of the scientific community. The AP task aims to reveal as much as possible demographic-information from a given set of authors [10]. For example, age, gender, native language, personality traits, cultural background, etc. The AP task has a wide applicability and could have a broad impact in a number of problems. For instance, in forensics the profile of authors could be used as valuable additional evidence, and in marketing the on-line reviews of companies/products could be exploited to improve targeted advertising.

The AP task at PAN17 is focused on the recognition of gender and language variety [21]. According to the literature, these AP tasks have been approached by several researchers [24,4]. Most of these efforts have been devoted to the analysis of the textual representation and features (e.g., words, POS Tags, etc.) [24,17,15,23]. Regardless of

the novel textual features and representations, most of them fail in capturing accurate information from informal documents. Especially in the social media domain, where the easiness of writing/sending messages leads people to make many grammatical and spelling errors. The previous situation captured the attention of some researchers, who began to model high level aspects (e.g., semantical and structural information) for the AP task [14,2,23]. Nevertheless, these proposals are not necessarily aligned with the main objective of AP; to model *groups of authors*. According to the literature of AP task in social media, the methods that build high level features based on relationships among groups of authors have been useful for boosting the performance [22,23]. In this paper we propose to study this aspect of groups of users. This is, instead of extracting coarse levels of analysis (groups of users), we propose to extract finer levels of granularity (by observing single users). For this we propose the idea of representing user by relationships with other users. This natural extension provides fine grained semantic representation of users by exploiting relationships with other individuals in the dataset. For example in language variety identification, a user could be known by exposing its relationship with their compatriots and non-compatriots. Our experimental evaluation shows that this approach improves even more the representation of documents in the AP task, and also mitigates the common problems of other standard representations (e.g. the Bag-of-Terms, BoT), for example: i) high dimensionality, and ii) the sparseness of the representation. Experimental results using the latter ideas also seem promising and competitive compared to other approaches such as Word2Vec in PAN 2017 collections. To the best of our knowledge, there are no reported results on AP using a similar strategy.

The remainder of this paper is organized as follows: Section 3 introduces the proposed approach. Section 2 presents some of the related work of this paper. Section 4 and 5 explains the datasets and evaluation methodology for this proposal. Finally Section 6 outlines the main conclusions and future avenues of inquiry.

2 Related Work

In this section we review the AP related work from the computational linguistics perspective. According to the literature, a wide range of different approaches have been proposed for the AP task. The different methods for learning specific textual patterns range from simple lexical approaches to elaborated strategies requiring syntactic/semantic analysis of the documents [10,5,8]. Notwithstanding the usefulness of these features, most of them are only relevant in domains with formal documents (i.e., books, articles, etc.). In the case of social media, the majority of the works have focused on using *content* and *stylistic* features [20,19,12,18]. In this direction, several works have suggested that content words are usually much more relevant than style features. For example, an analysis of information gain presented in [24], showed that the most relevant attributes for gender prediction are those related with content words, for example: *linux* and *office* for discriminating males, whereas *love* and *shopping* for discriminating females. Furthermore, Schler et al. also concluded that syntactic features are less useful than very basic lexical thematic features when analysing blogs [24]. Other works have also considered interesting stylistic features, namely slang vocabulary and the average

sentence length, but in all the cases these features have been used in combination –as a complement– of content features [3,9]. In this work, we use the well-known lexical features (content and style) in order to feed a representation that captures relationships among user-documents.

In all previous works authors have proposed interesting strategies to exploit the different aspects of the AP task. However, most of those works have only marginally explored the finer details of the high diversity in groups and subgroups of authors [22,23]. In this regard, the Second Order Attributes (SOA) representation has been one of the most notable works for AP in social media [23,1,14]. The key idea of SOA is closely related to the main objective of the Author Profiling; to model groups of authors. For example, the first version of the SOA [13] pushes to the limit this idea by building very low-dimensional document vectors. The key element was to build one feature per target profile [1]. Thus, the SOA assumed that there exists certain homogeneity among all documents/user-documents that belong to a same class (profile). For this reason, the second version of SOA [12,14] introduces novel improvements to capture finer details of the high diversity in subgroups that make each target profile unique [2]. The latter works have made evident the relevance of exploiting the existing knowledge in the dataset, therefore it is promising to explore novel alternatives in this direction.

In this work, we attempt to evaluate the relevance of user-documents relationships for the task of AP in social media. Our main hypothesis is that user-documents in the same class (profile), should have similar preferences of topics (preferences for certain topics), and therefore they are the cornerstone to reveal profiling cues in social media domains. More specifically, in this work we are proposing to exploit the relationship of each user with other users (finer level of granularity), instead of relationships with groups of users (coarse level of granularity). In particular, we propose to model these features by capturing the distribution of word-usage among all user-documents, in order to automatically extract the relationships from the given user-documents collection.

3 Methodology

This section presents the general framework for the User Specific Representation (USR). The USR follows the ideas from the document occurrence representation in [11], in order to achieve relationships between the target user-documents and other user-documents. By using USR, we aim to represent each user-document in an enriched distributional space that highlights relevant information with other user-documents. The key idea is to exploit the hypothesis that words occurring in similar user-documents should have similar representations, and therefore are useful to characterize the relationships among users [14]. In this way, USR requires of two steps. In the first step, terms (e.g., words) are represented in a new *user-document* space. In the second step, the documents (user-documents) are mapped to the new *user-documents* space. This mapping is done by

[1] This means that in a gender classification scenario (e.g., female vs male), documents would be represented using only 2 features.

[2] For example, while the largest group of males writes about sports and technology, there are other small groups of males interested in stereotypically female topics such as family and friends.

aggregating the representation of the terms that occur in the instance. In the following sections we describe in detail how the USR computes relationship values using words as terms.

3.1 User Specific Representation (USR)

Let $\mathcal{U} = \{(U_1, y_1), \ldots, (U_n, y_n)\}$ be a training set of labeled user-documents, that is, $\mathcal{U}$ is a collection of $n-$tuples of user-documents (U_i) and category-labels (y_i). Also let $\mathcal{V} = \{v_1, \ldots, v_m\}$ denote the vocabulary of terms (e.g., words). The core idea of USR consists in capturing the semantics of a word by observing the distribution of occurrence statistics over the user-documents in the dataset [11]. More formally, each word v_i is represented as a vector $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,|\mathcal{U}|} \rangle$, where $|\mathcal{U}|$ is the number of users-documents in the training collection, and $t_{i,k}$ indicates the relevance of the user-document U_k to characterize v_i. Equation 1 presents the above ideas.

$$t_{i,k} = \mathit{df}(v_i, U_k) \cdot \log \frac{|\mathcal{U}|}{|\mathcal{N}_k|} \tag{1}$$

where $\mathcal{N}_k \subseteq \mathcal{V}$ is the set of different terms in the user-document U_k, and $\mathit{df}(v_i, U_k)$ is defined in Equation 2.

$$\mathit{df}(v_i, U_k) = \begin{cases} 1 + \log(\#(v_i, U_k)) & \text{if } \#(v_i, U_k) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\#(v_i, U_k)$ indicates the frequency of term v_i in U_k. The intuitive idea is that, the importance of a user-document U_k, is given by the frequency of the term v_i in U_k. Also note that the number of different terms contained in U_k is inversely proportional to its contribution to represent v_i. Note that, the distributional representation of each term $\mathbf{t_i}$ is normalized so that $||\mathbf{t_i}||_2 = 1$.

Finally, the user-document representation is obtained by aggregating the representation of terms that occur in the user-document. This is $\mathbf{U_k} = \sum_{v_i \in U_k} \mathbf{t_i}$, where $\mathbf{t_i}$ denotes the representation of the term v_i. One should note that, the relevance of word v_i in the user-document U_k is given by the frequency of the term in the document under analysis.

4 Corpora and Experimental Framework

The PAN 2017 corpora is composed by 4 collections in different languages (English, Spanish, Portuguese and Arabic). For each dataset, there are labels of gender (male, female) and language-varieties. We approached the PAN 2017 AP task as a classification problem. For this we build two separate classification models: variety and gender models. For the evaluation, in all experiments we use the following experimental configurations in the training dataset:

1. **Basic Textual Features**[3]: For English, Spanish, and Portuguese we basically used as terms: words, contractions, words with hyphens, punctuation marks and a set of common slang vocabulary. For the Arabic language we also used a straightforward preprocessing; we split sentences into tokens using the blank space.
2. **Number of Features**: For all the training collections we used the top frequent 15k terms as features. We determine this by an empirical evaluation testing values from 5k to 20k features[4]. This preliminary experimental evaluation is useful to determine an appropriated number of Features [14].

4.1 Experimental Framework

The aim of this evaluation is to compare the performance of the proposed USR and relevant methods in the AP for social media. For this purpose we separately evaluate the traditional Bag-of-Terms (BoT) using TF-IDF as the weighing scheme. The BoT has been a strong baseline in many AP social media domains [14]. In the experiments, we also show results of a word embedding approach, which is based in the well-known Word2Vec (W2V) [16]. This approach trains W2V representation of the words by using and end-to-end training on each dataset, then builds a document representation averaging the word-vectors of the words contained in the target user-document. Finally we consider the Second Order Attributes and the Subprofile Specific Representations (SSR) approaches [13,12,14], which to some extent, have been present in the winning approaches of all previous PAN AP editions [23]. For the classification step, we firstly build the representation of terms and documents using the latter methodologies. Then we evaluate a standard LibLINEAR classifier without any parameter optimization [7], by using a 10 Fold Cross Validation (FCV) framework.

5 Evaluation

In these experiments we are interested in exploring the contribution of user-document relationships (captured by USR) in the AP task. The target problems are language variety and gender prediction. In addition to the USR results, we also present results from several methods reported in the literature of AP for social media.

From Table 1 and 2 it can be seen that the proposed approach (USR) outperforms BoT by a considerable difference. This is interesting, since the BoT has shown outstanding performance in many different AP tasks [22,23,14]. Similarly, there is also an improvement compared to the SOA based approaches (SOA [13] and SSR [12,14]), although the difference is smaller, especially compared with SSR [5]. We hypothesize that this is because of the finer-level of granularity in the USR representation, which provides specific relationship values for all user-documents. The latter can be seen as

[3] In this work we use well-known basic features for AP since we are interested in the high level instance (user-document) representation.

[4] In general, for our tested representations in PAN17 datasets, taking more than 15k features does not have any significant improvement in the classification performance

[5] In this version of SSR we generated 50 subprofiles (we tested values from 10 to 100) for each target profile (see [14] for more details about subprofile generation).

Table 1. Experimental evaluation for variety prediction. Results show the accuracy performance of the proposed User Based Representation (USR), and the main approaches in the AP literature.

Accuracy for Variety prediction in PAN17 copora					
Dataset	Representation	English	Spanish	Portuguese	Arabic
Train	BoT	76.1	89.4	98.3	77.3
	W2V	63.4	81.5	98.2	64.4
	SOA	75.7	87.2	98.6	75.6
	SSR	84.8	92.9	98.7	80.8
	USR	**88.3**	**94.2**	**98.7**	**82.3**
Test	USR	85.67	94.32	98.25	81.19

Table 2. Experimental evaluation for gender prediction. Results show the accuracy performance of the proposed User Based Representation (USR), and the main approaches in the AP literature.

Accuracy for Gender prediction in PAN17 copora					
Dataset	Representation	English	Spanish	Portuguese	Arabic
Train	BoT	78.9	72.5	80.2	76.6
	W2V	77.4	73.6	74.8	73.9
	SOA	77.5	68.4	73.2	74.2
	SSR	82.1	74.7	80.8	77.1
	USR	**82.3**	**78.3**	**83.8**	**79.3**
Test	USR	81.71	80.14	82.38	77.63

a weakness or a strength depending on the domain and the nature of social media documents. SSR summarizes in some way the information in the collection by building groups and subgroups, therefore less features are needed to represent the document. On the other hand, USR provides a high level of detail that might be not desired in specific scenarios where the required computational cost is crucial. Finally Word2Vec-based representation obtained low performance, this could be due to the end-to-end training in the data collection, where more data could be necessary to build a better model. We individually validate USR using the Wilcoxon Signed Rank [6] test against: BoT, W2V and SOA. The output obtained by this test is above of 98% of statistical confidence.

We believe that the performance in USR is because finding user-document relationships in the target classes (profiles), provides a more detailed perspective for documents. In this regard, USR is a novel representation in AP task for social media that capture details at a finer level of user-document granularity never seen before for AP task. Thus, the approach presented in this paper is an effective alternative to address the AP task in different social media domains, where documents present challenging difficulties hindered the accurate work of most natural language processing tools.

6 Conclusions

In this paper we presented a novel idea to approach the AP task. The proposal extracts relationship values from user-documents in the dataset, in order to improve the pre-

diction of the unknown user profiles. For this, we exploit the distributional hypothesis by means of a document occurrence representation, which in the context of AP is a User Specific Representation (USR). Such USR models target user-documents by computing specific relationship values with the other user-documents. The intuitive idea is that each user should have high relationship values with users of its own profile. For example, in language variety identification the relationship values of a Mexican user-document should be high compared with other Mexicans, but low compared with people from other countries.

To the best of our knowledge, this is the first time that AP is addressed using this kind of user-document relationships. The latter relationships help to improve the classification performance in most of the cases. Using these distributional attributes, the classifier can keep good classification rates. This is due to the relationship among terms and user-documents, which provides few but more detailed predictive attributes. We have shown better experimental results than the standard BOT, Word2Vec and SOA, which has shown to be useful in all PAN editions.

References

1. Alvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villasenor-Pineda, L., Escalante, H.J.: Inaoe's participation at pan'15: Author profiling task. In: Working Notes Papers of the CLEF (2015)
2. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Meza, I.: Evaluating topic-based representations for author profiling in social media. In: Ibero-American Conference on Artificial Intelligence. pp. 151–162. Springer (2016)
3. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday 12(9) (2007)
4. 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)
5. Bergsma, S., Post, M., Yarowsky, D.: Stylometric analysis of scientific articles. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 327–337. Association for Computational Linguistics (2012)
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30 (2006)
7. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
8. Garera, N., Yarowsky, D.: Modeling latent biographic attributes in conversational genres. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. vol. 2, pp. 710–718. Association for Computational Linguistics (2009)
9. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers age and gender. In: Third International AAAI Conference on Weblogs and Social Media (2009)
10. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
11. Lavelli, A., Sebastiani, F., Zanoli, R.: Distributional term representations: an experimental comparison. In: CIKM. pp. 615–624. ACM (2004)
12. López-Monroy, A.P., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L.: Using intra-profile information for author profiling. In: CLEF (Working Notes) (2014)

13. López-Monroy, A.P., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Villatoro-Tello, E.: INAOE's participation at PAN'13: Author profiling task. In: CLEF (Working Notes) (2013)
14. 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)
15. López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A new document author representation for authorship attribution. In: Pattern Recognition: 4th Mexican Conference, MCPR 2012, Huatulco, Mexico, June 27-30, 2012. Proceedings. pp. 283–292. Springer (2012)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
17. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: How old do you think i am?: A study of language and age in twitter. In: Seventh International AAAI Conference on Weblogs and Social Media (2013)
18. Ortega-Mendoza, R.M., Franco-Arcega, A., López-Monroy, A.P., Montes-y Gómez, M.: I, me, mine: The role of personal phrases in author profiling. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 110–122. Springer (2016)
19. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the author profiling task at PAN 2014. In: CLEF (Online Working Notes/Labs/Workshop). pp. 898–927 (2014)
20. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Valencia, Spain, September. pp. 23–26 (2013)
21. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
22. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF. sn (2015)
23. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF (2016)
24. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. pp. 199–205 (2006)