INAOE's participation at PAN'15: Author Profiling task

Notebook for PAN at CLEF 2015

Miguel Ángel Álvarez-Carmona, A. Pastor López-Monroy,
Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, and Hugo Jair-Escalante

Language Technologies Laboratory, Department of Computer Science,
Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro No. 1, C.P. 72840, Pue. Puebla, México
{miguelangel.alvarezcarmona,pastor, mmontesg, villasen, hugojair}@inaoep.mx

Abstract In this paper, we describe the participation of the Language Technologies Lab of INAOE at PAN 2015. According to the Author Profiling (AP) literature, the most useful textual attributes correspond to stylistic and thematic features. In this paper we take such stylistic and thematic information into a new higher level exploiting a combination of *discriminative-stylistic* and *descriptive-thematic* representations. For this we use dimensionality reduction techniques on the top of typical stylistic and thematic textual features for AP task. The main idea is that each representation, using the full feature space, automatically highlights the different stylistic and thematic properties in the documents. Specifically, we propose the joint use of Second Order Attributes (SOA) and Latent Semantic Analysis (LSA) techniques to highlight *discriminative-stylistic* and *descriptive-thematic* properties respectively. In order to evaluate our approach, we compare the proposed approach against a standard Bag-of-Words (BOW), SOA and LSA representations using the PAN 2015 corpus for AP. Experimental results in AP show that the combination of SOA and LSA outperforms the BOW and each individual representation, which gives evidence of its usefulness to predict gender, age and personality profiles. More importantly, according to the PAN 2015 evaluation, the proposed approach are in the top 3 positions in every dataset.

1 Introduction

The Author Profiling (AP) task consists in knowing as much as possible about an unknown author, just by analysing a given text [4]. The interest in AP tasks has captured the attention of the scientific community in recent years. This is due, in part, to the potential of the huge amount of the user-generated textual information on the internet. In this context, several applications related to AP are emerging, some of them have to do

with e-commerce, computer forensics and security. There are several ways to address the AP task. One of them is to approach it as a single-label multiclass classification problem, where the target specific profiles (e.g., male, female) represent the classes to discriminate.

Broadly speaking, in text classification tasks there are three general-key procedures; i) the extraction of textual features, ii) the representation of the documents, and iii) the application of a learning learning algorithm. In the context of the AP tasks, for the first step, specific lexical (e.g., simple words, function words) [4] and syntactical features (e.g., POS tags) [14] have proven to be highly discriminative for some specific profiles. Regarding to the last two steps, the representation of documents and the learning algorithm are the most common-effective approaches for AP tasks consist in using the Bag-of-Words formulation (e.g., histograms of the presence/absence of textual features) [15] and Support Vector Machines [2][5] respectively.

According to the AP task literature, most of the work has been devoted to the first step: to identify the most useful-interesting textual features for the target profiles. In spite of the usefulness of previous interesting textual features and the good results achieved by BoW-SVM, the research community has put little effort to deepen in the second and third steps: alternative representations and learning algorithms for the AP task. The main shortcomings of the BoW-SVM approach are well known from other text mining task. The most relevant ones for the AP task are:

- **The order of words are lost:** Once the BoW is built, terms are represented as histograms of the occurrence of textual features, which loses useful information about the context of the words.
- **High dimensionality and sparseness:** The noise presented in social media documents stressed those characteristics, which depending of the learning method, could affect in different ways the general approach. For example, the effectiveness, the required training time or the interpretation of the representations.

To overcome the latter shortcomings, in this paper we focus in the second step in order to improve the representation of tweets. The main goal of our approach is to compute high quality *stylistic* and *thematic* features built on the top of the state-of-the-art typical textual features (e.g., content words, function words, punctuation marks, etc.). For this, we propose to combine two state-of-the-art dimensionality reduction techniques that best contribute to automatically stress the contribution of the *stylistic* and *thematic* textual features. According to the literature the most frequent textual features (e.g., function words, stopwords, punctuation marks) provide important clues about the style of the authors. For this we need a representation highly based in term frequencies,

that stresses the contribution of such stylistic attributes and produces highly *discriminative* document representations. To capture stylistic information contained among textual features we use Second Order Attributes (SOA) computed as in [8]. On the other hand, relevant thematic information usually are in *descriptive* terms, terms that are frequent only in some specific documents or classes. In this way, to represent documents we bring ideas from the information retrieval field exploiting the Latent Semantic Analysis (LSA) [16]. LSA represents terms and documents into a new semantic space. This is done performing a singular value decomposition using a Term Frequency Inverse Document Frequency (TFIDF) matrix. The descriptive terms and documents representation are stressed under the LSA formulation throwing out the noise, but emphasizing strong patterns and trends. To the best of our knowledge, the idea of representing documents using the combination *stylistic-discriminative* and the *thematic-descriptive* high-level features through dimensionality reduction techniques have never been explored before in AP task. Thus, it is promising to bring together two of the best document representations to better improve the AP; that is precisely the propose of this work.

The rest of this paper is organized as follows: Section 2 introduces the proposed representation, in Section 3 some characteristics of the corpus PAN15 are explained briefly, Section 4 explains how we performed the experiments and the results we obtained, finally Section 5 shows our conclusions.

2 Exploiting Discriminative-Stylistic and Descriptive-Thematic features

Along this section we briefly describe each representation and the proposed strategy to compute the final representation of documents. In Section 2.1 we explain the SOA representation to get the discriminative-stylistic features. In Section 2.2 we explain the LSA algorithm with which we intend to get descriptive-thematic features. Finally, in Section 2.3 we explain how we join these representations for the AP task.

2.1 Computing Discriminative-Stylistic Features

The stylistic textual features have proven to be useful for AP task [11]. A plenty of the style textual attributes in text mining tasks (e.g., Author Profiling, Authorship Attribution, Plagiarism Detection) have been associated with highly frequent terms [12]. For example, observing the frequency of stopwords and punctuation marks exposes clues about the author of a document. In gender identification observing the distribution of

specific function words and determiners have proven to be also useful [11]. Second Order Attributes (SOA) proposed in [8] is a supervised frequency based approach to build document vectors in a space of the target profiles. Under this representation, each value in the document vector represents the relationship of each document with each target profile.

The representation as described in [8] has two keys steps. i) To build words vectors in an space of profiles and ii) to build documents vectors in an space of profiles. In the former step, for each vocabulary term t_j, a $\mathbf{t_j} = \langle tp_{1j}, ..., tp_{mj} \rangle$ vector is computed. Where each tp_{mj} is a frequency-based-value that represents the relationship between term t_j and the profile p_m. In the latter step the representation of documents is built using a weighted by frequency aggregation of the term vectors contained in the document (see Equation 1).

$$\mathbf{d_k} = \sum_{t_j \in D_k} \frac{tf_{kj}}{lenght(d_k)} \mathbf{t_j} \tag{1}$$

where D_k is the set of terms that belongs to document d_k.

For more details please refer to [8].

2.2 Computing Descriptive-Thematic Features

Besides the usefulness of stylistic features, thematic information has proven to be an important aspect for the AP Task [11]. For example, several works have shown evidence that groups of people of the same age and gender write generality about the same topics. For this reason we exploit the Latent Semantic Analysis (LSA). LSA is a technique that can associate words and it contribution to automatically generated concepts (topics) [1]. This is usually named the latent space, where documents and terms are projected to produce a reduced topic based representation. We hypothesises that under the aforementioned latent space, we can better expose thematic relevant information for the AP task.

LSA is a method to extract and represent the meaning of the words and documents. LSA usually is built from a matrix $\mathbf{M}$ where m_{ij} is typically represented by the TFIDF [13] of the word i in document j. LSA uses the Singular Value Decomposition (SVD) to decompose $\mathbf{M}$ as follows.

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2}$$

Where The $\mathbf{\Sigma}$ values are called the singular values and $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors respectively. $\mathbf{U}$ and $\mathbf{V}$ contains a reduced dimensional representation

of words and documents respectively. $\mathbf{U}$ and $\mathbf{V}$ emphasizes the strongest relationships and throws away the noise [6]. In other words, it makes the best possible reconstruction of the $\mathbf{M}$ matrix with the less possible information [7]. Using $\mathbf{U}$ and $\mathbf{V}$ computed only from the training documents, words and documents are represented for training and test. For more details please refer to [16].

2.3 Exploiting the jointly use of Global-Local Semantic Features

The idea is to use the representations built under the whole feature space to automatically highlight the *discriminative-stylistic* and *descriptive-thematic* properties in documents. The intuitive idea is to take advantage of both approaches in a representation using early fusion. Let $\mathbf{x}_j$ be the $j-th$ training instance-profile under LSA representation with k dimensions and $\mathbf{y}_j$ be the same instance-profile under the SOA representation with m dimensions, the final representation is show in Expression 3.

$$\mathbf{z_j} = \langle x_{j1}, \ldots, x_{jk}, y_{j1}, \ldots, y_{jm} \rangle \tag{3}$$

The collection of training documents are finally represented as:

$$\mathbf{Z} = \bigcup_{d_j \in D} \langle \mathbf{z_j}, c_j \rangle \tag{4}$$

Where c_j is the class of the $j-th$ training instance-profile.

3 Data Collection

We have approached the PAN 2015 AP task as a classification problem. PAN 2015 corpora is composed by 4 datasets in different languages (Spanish, English, Italian and Dutch). Each dataset has labels of gender (male, female), age [1] (18-24, 25-34, 35-49, 50-xx) and five personality traits values (extroverted, stable, agreeable, conscientious, open) between -0.5 and 0.5. In Table 1 we show the number of Author-Profiles per language.

For personality identification Table 2 shows the relevant information (in terms of classes). For each language it shows the range and the number of the classes for each trait[2]. For personality we consider each trait value in the train corpus as a class. For

[1] Age data for Italian and Dutch languages are not available.

[2] The ranges with asterisk indicate that a value between the range is missing. For example, in Spanish (extroverted and conscientious) the -0.1 is missing.

example, if only two values (e.g., 0.2 and 0.3) are observed in the training corpus, then we built a two class classifier (e.g., 0.2 and 0.3) [3].

Table 1. Description of the dataset

Language	Author-Profiles
English	152
Spanish	100
Italian	38
Dutch	34

Table 2. The personality traits information by language

	English		Spanish		Italian		Dutch	
Trait	Range	Classes	Range	Classes	Range	Classes	Range	Classes
Extroverted	[-0.3,0.5]	9	[-0.3,0.5]*	8	[0.0,0.5]*	5	[0.0,0.5]	6
Stable	[-0.3,0.5]	9	[-0.3,0.5]	9	[-0.1,0.5]	7	[-0.2,0.5]	8
Agreeable	[-0.3,0.5]	9	[-0.2,0.5]	8	[-0.1,0.5]*	6	[-0.1,0.4]	6
Conscientious	[-0.2,0.5]	8	[-0.2,0.5]*	7	[0.0,0.4]	5	[-0.1,0.4]	6
Open	[-0.1,0.5]	7	[-0.1,0.5]	7	[0.0,054]	6	[0.1,0.5]	5

4 Experimental Evaluation

4.1 Experimental Settings

We use for each experiment the following configuration: i) for terms we use words, contractions, words with hyphens, punctuation marks and a set of common emoticons, ii) we consider the terms with at least 5 occurrences in the corpus, iii) the number of concepts for LSA is set to $k = 100$. We perform an stratified 10 cross fold validation (CFV) using the training PAN15 corpus and a LibLINEAR classifier [3]. In order to determine the full profile of a document (gender, age and the 5 personality traits) we built one classifier to predict each target profile for each language.

[3] For each personality trait in each language the number of the classes are variables between them, see Table 2

4.2 Experimental Results

The aim of this first experiment is to analyse the performance of LSA, SOA and the BOW approach in the AP tasks. We experiment with LSA and SOA separately and finally with the two approaches together. We are interested in observing the contribution of discriminative-stylistic (captured by SOA) and descriptive-thematic (captured by LSA) information in the AP task. For gender prediction, in Table 3 we can see that considering the individual representations, LSA obtains the best results, which outperforms the BOW approach in every language. When LSA and SOA are together the result only improves in English, which is an important remark since the English language is the bigger-robust collection (see Table 1). The following conclusions can be outlined from Table 3:

Table 3. Detailed classification accuracy to gender

Language	BOW	SOA	LSA	LSA+SOA
English	74.00	70.86	74.34	**78.28**
Spanish	84.00	74.00	**91.00**	**91.00**
Italian	76.31	73.68	**86.84**	**86.84**
Dutch	82.35	91.07	**91.17**	**91.17**

- The descriptive-thematic information captured by LSA is the most relevant information for gender prediction in PAN 2015 AP dataset. This is because LSA obtained the best average individual performance.
- The pure discriminative-stylistic captured by SOA only outperforms BOW in Dutch documents. But the combination of LSA and SOA obtained an improvement of around 4% in accuracy for English gender detection. We think, SOA could improve the results if more documents are available [4].

For age prediction Table 4 shows the experimental results. Recall that the age data is available only for English and Spanish languages. As in the last experiment LSA obtains the best individual performance, but in this experiment the combination of LSA and SOA obtains an improvement in both collections. It is worth noting that despite of the small datasets, for age prediction SOA could contribute to improve the classification[5].

[4] SOA has proven outstanding results in recent years in the PAN AP tracks [10,9].
[5] The best results for SOA in previous PAN AP editions have been for age prediction

Table 4. Detailed classification accuracy to age

Language	BOW	SOA	LSA	LSA+SOA
English	74.83	68.21	78.94	**79.60**
Spanish	80.00	74.00	81.00	**82.00**

Finally for personality prediction Table 5 shows the performance of BOW and LSA plus SOA performance by language in the personality detection task. For this experiment, although the results seems promising they should be taken with caution. This is due to the lack of data and the number of classes that we consider (one class for each observed value) one correct/wrong predicted instance is enough to change the results considerably. For this specific experiment in personality, we built a representation on the entire dataset, then we evaluate using a 10CFV. In general, the results suggest that the combination of LSA plus SOA gets similar or better results than the typical BOW approach. Given evidence of the usefulness of the *discriminative-stylistic* features and the *descriptive-thematic* features.

Table 5. Detailed classification accuracy for personality

	English		Spanish		Italian		Dutch	
Trait	BOW	LSA+SOA	BOW	LSA+SOA	BOW	LSA+SOA	BOW	LSA+SOA
Extroverted	64	**87**	62	**87**	65	**94**	64	**91**
Stable	56	**85**	69	**91**	52	**94**	61	**94**
Agreeable	60	**80**	62	**84**	71	**92**	61	**88**
Conscientious	61	**78**	62	**86**	57	**94**	67	**91**
Open	65	**86**	62	**74**	55	**84**	64	**97**

5 Conclusions

In this paper, we have explored a new combination of document representations for AP task. The main aim of this work was to experiment the with *descriptive-thematic* (LSA) and *discriminative-stylistic* (SOA) features. We found that the *descriptive-thematic* information is very useful, which confirms several findings in the literature. Moreover, we also find that *discriminative-stylistic* information could improve the results when it is combined with *descriptive-thematic*. This indicates that LSA captures very important information which in turn can be complemented with the SOA stylistic information.

Acknowledgment Álvarez-Carmona thanks for doctoral scholarship CONACyT-Mexico 401887.

References

1. Dumais, S.T.: Latent semantic analysis. Annual review of information science and technology 38(1), 188–230 (2004)
2. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Author profiling for english emails. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07). pp. 263–272 (2007)
3. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
4. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
5. Koppel, M., Schler, J., Zigdon, K.: Determining an author's native language by mining a text for errors. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 624–628. ACM (2005)
6. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes 25(2-3), 259–284 (1998)
7. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of latent semantic analysis. Psychology Press (2013)
8. Lopez-Monroy, A.P., Gomez, M.M.y., Escalante, H.J., Villasenor-Pineda, L., Villatoro-Tello, E.: Inaoe's participation at pan'13: Author profiling task. In: CLEF 2013 Evaluation Labs and Workshop (2013)
9. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In: Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes) (2014)
10. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. Notebook Papers of CLEF pp. 23–26 (2013)
11. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6, pp. 199–205 (2006)
12. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538–556 (2009)
13. Turney, P.: Mining the web for synonyms: Pmi-ir versus lsa on toefl (2001)
14. Van Halteren, H.: Linguistic profiling for author recognition and verification. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. p. 199. Association for Computational Linguistics (2004)
15. Wang, G., Hoiem, D., Forsyth, D.: Building text features for object image classification. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1367–1374. IEEE (2009)

16. Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A.: Latent semantic analysis. In: Proceedings of the 16th international joint conference on Artificial intelligence. pp. 1–14. Citeseer (2004)