Information Retrieval Features for Personality Traits

Edson Roberto Duarte Weren

vveren@gmail.com

Abstract. This paper describes the methods employed to solve the Author Profiling task at PAN-2015. The main goal was to test the use of features derived from Information Retrieval to identify the personality traits of the author of a given text. This paper describes the features, the classification algorithms employed, and how the experiments were run. Also, I provide a comparative analysis of my results compared to those of other groups.

Keywords: Information Storage and Retrieval, Document and Text Processing.

1 Introduction

Author Profiling, which has a growing importance in applications in forensics, marketing, and security[1], deals with the problem of finding as much information as possible about an author, just by analyzing an text produced by that author.

This paper reports on the my participation at the third edition of the Author Profiling task, organized in the scope of the PAN Workshop series, which is collocated with CLEF2015. More details about the task and the workshop can be found in the overview paper [3]. The task requires that participating teams come up with approaches that take a text as input and predict the gender (male/female), the age group (18-24, 25-34, 35-49, or 50+) and the author's personality traits (extroverted, stable, agreeable, conscientious, or open in a range from -0.5 to 0.5).

2 Identifying Author Profiles

The underlying assumption was that authors from the same gender, age group or personality traits tend to use similar terms and that the distribution of these terms would be different across genders, age groups and personality traits. To implement this notion, all conversations were indexed using an Information Retrieval engine and then, the conversation to be classified was treated as a query. The idea is that the conversations retrieved (*i.e.*, the most similar to the query) are the ones from the same gender, age group, and personality traits.

The training dataset was composed of conversations (XML files) about various topics grouped by author. Conversations were in English, Spanish, Italian, and Dutch and were annotated with the gender, age group, and the personality traits of the author. A complete description of the dataset may be found in `http://pan.webis.de/`.

2.1 Features

The texts from each author, or the documents, were represented by a set of 288 features (or attributes).

The complete set of texts was indexed by an Information Retrieval (IR) System in a manner similar to that used in [4–6]. Then, the text that to be classified was used as a query and the k most similar texts were retrieved. The ranking is given by the Cosine or Okapi metrics as explained below.

Cosine These features are computed as an aggregation function over the top-k results for each age, gender, and personality trait obtained in response to a query composed by the keywords in the text to be classified. Three types of aggregation functions were tested, namely: count, sum, and average. For this featureset, queries and documents were compared using the cosine similarity (Eq. 1). For example, if we retrieve 10 documents in response to a query composed by the keywords in q, and 5 of the retrieved documents were in the 18-24 age group, then the value for 18-24_cosine_avg is the the average of the 5 cosine scores for this class. Similarly, 18-24_cosine_sum is the summation of such scores, and 18-24_cosine_count simply counts how many retrieved documents fall into the 18-24_cosine_count category.

$$COSINE = (c, s)\frac{\boldsymbol{c.q}}{|\boldsymbol{c}||\boldsymbol{q}|} \tag{1}$$

where $\overrightarrow{c}$ and $\overrightarrow{q}$ are the vectors for the document and the query, respectively. The vectors are composed of $tf_{i,c} \times idf_i$ weights where $tf_{i,c}$ is the frequency of term i in document c, and $IDF_i = \log \frac{N}{n(i)}$ where N is the total number of documents in the collection, and $n(i)$ is the number of documents containing i.

Okapi Similar to the previous, these features compute an aggregation function (average, sum, and count) over the retrieved results from each gender, age, and personality traits group that appeared in the top-k ranks for the query composed by the keywords in the document. For this featureset, queries and documents were compared using the Okapi BM25 score (Eq. 2).

$$BM25(c, q) = \sum_{i=1}^{n} IDF_i \frac{tf_{i,c} \cdot (k_1 + 1)}{tf_{i,c} + k_1(1 - b + b\frac{|D|}{avgdl})} \tag{2}$$

where $tf_{i,c}$ and IDF_i are as in Eq. 1 $|d|$ is the length (in words) of document c, $avgdl$ is the average document length in the collection, k_1 and b are parameters

that tune the importance of the presence of each term in the query and the length of the text. In my experiments, i used $k_1 = 1.2$ and $b = 0.75$.

2.2 Experiments

The steps taken to process the datasets and run the experiments were the following:

1. Pre-process the conversation in the training data to tokenize (only during testing, stemming and stopword removal was performed but without significant gains).
2. Use each conversation as queries.
3. Index 100% of the pre-processed conversations with a retrieval engine. Zettair[1], which is a compact and fast search engine developed by RMIT University (Australia), was used for indexing and querying. Zettair implements several methods for ranking documents in response to queries and calculates cosine and Okapi BM25.
4. Compute the features using the results from the queries submitted to Zettair. The top-10 scoring conversations were retrieved.
5. Train the classifiers and generate the models. Weka [2] was used to build the classification models.
6. Use the trained classifiers to predict the classes of the conversations used as queries. Once the classifiers are trained, they can be used to predict the classes for new unlabelled conversations. Thus, the conversations from the test data were treated as queries and went through steps 1, 4 and 6.

2.3 Training the Classifiers

Twenty-six classifiers are necessary, since there are four languages and seven dimensions in each (age [only English and Spanish], gender, and personality traits [extroverted, stable, agreeable, conscientious, and open]). All results in this section refer to experiments run on the *training data* only. The predictions of the classifiers were compared two settings (*i*) using all 288 features an (*ii*) using just a subset of 6 to 22 features (produced by BestFirst subset evaluator).

Figure 1 shows the results comparing the accuracy of the runs that use all features and the runs that use just the subset. Using the subsets had advantages in nearly all cases. The only exception was for age. This can be confirmed by Table 1, which shows results of paired t-tests that assess the significance of the difference between the runs that use all features and the runs that use a subset only. For the majority of the learning algorithms, the results of all runs are very close, with a slight advantage in favor of the runs with the selected subset of attributes.

Figure 2 shows the most accurate classifiers on the training data grouped by language. We can see that some languages had better performance than

[1] http://www.seg.rmit.edu.au/zettair/

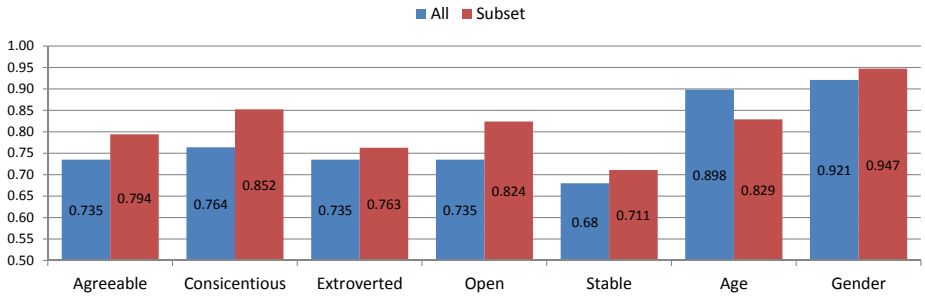


Fig. 1. Accuracy considering all and a subset of features

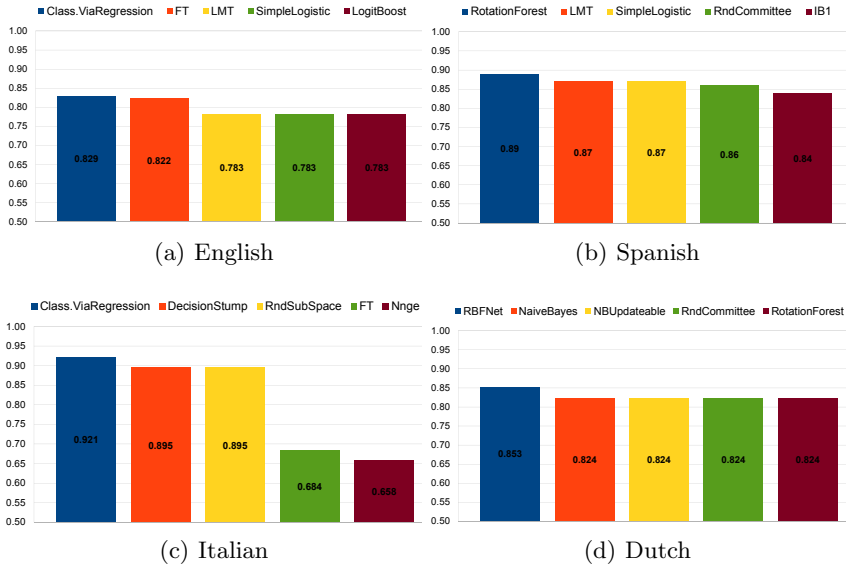
Table 1. Significance differences between pairs all/subset of features for four languages and seven dimensions

Dimensions	Language	P-value	Significant	Advantage of Use of the Features
Agreeable	EN	0.00	Yes	All
	ES	0.03	Yes	Subset
	IT-NL	0.00	Yes	Subset
Conscientious	EN	0.30	No	-
	ES	0.40	No	-
	IT-NL	0.16	No	-
	NL	0.04	Yes	Subset
Extroverted	EN	0.00	Yes	All
	IT-NL	0.07	No	-
	NL-ES	0.00	Yes	Subset
Gender	ES	0.01	Yes	Subset
	EN-IT-NL	0.00	Yes	Subset
Age	EN	0.00	Yes	Subset
	ES	0.07	No	-
Open	All	0.00	Yes	Subset
Stable	ES	0.06	No	-
	EN-IT-NL	0.00	Yes	Subset

others. While Italian had the best scores, English had the lowest. This difference could be explained by the fact that Italian has a more diverse morphology and a greater vocabulary compared to English and this may provide the classifier with more distinctive features. Regarding the choice of classifier, we can see that different languages had different classifiers as best performers. Classification via Regression, Random Committee, and Rotation Forest were among the top 5 in two cases each.

Figure 3 shows the most accurate classifiers for Age and Gender prediction. For age, we notice that a number of algorithms achieved similar results (around 0.8). For gender, RBFNetwork was the best.

Figure 4 shows the best classifiers for modelling personality traits. The Multilayer Perceptron is among the top performers in three out of five traits. The

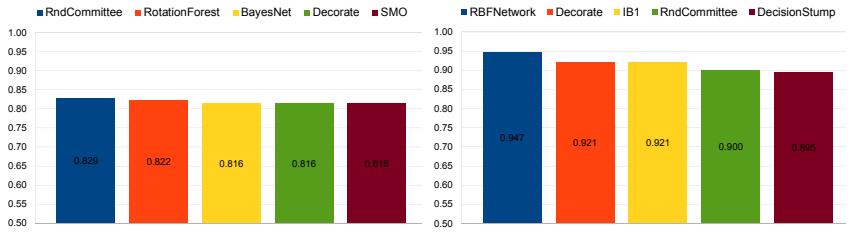


(a) English (b) Spanish

(c) Italian (d) Dutch

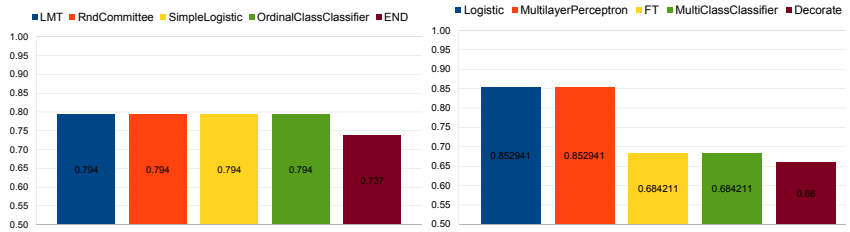
Fig. 2. Best classifiers based on Accuracy by Language

most notable cases in which there were large differences in the accuracies of the classifiers were for Conscientious and Open.

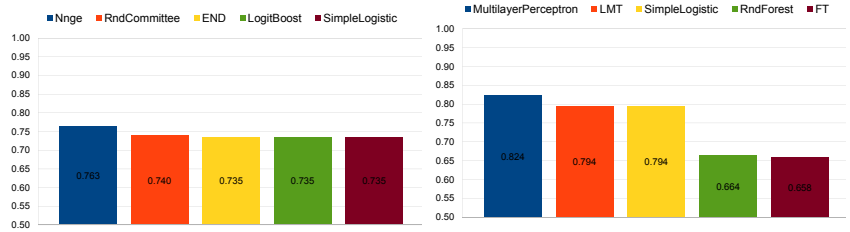


(a) Age (b) Gender

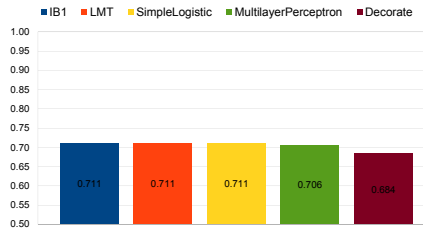
Fig. 3. Best classifiers based on Accuracy by Age and Gender



(a) Agreeable (b) Conscientious



(c) Extroverted (d) Open



(e) Stable

Fig. 4. Best classifiers based on Accuracy by Trait

3 Official Experiments

A pairwise comparison of the accuracies of the twenty-two teams which participated on the Author Profiling task was held, considering age, gender, personality traits and language. In this work, for the systems to be significantly different from each other, they had to have $p < 0.05$.

As a result, system proposed in study is not significantly different from systems that scored best, considering as an example the small set of training data used: English, Spanish, Italian, and Dutch - 152, 100, 38 and 34 files, respectively.

Comparing the results on the training and test datasets, a drop of about ten percentage points was observed. Overall results on the training data were 0.8171 and on the test data the final score was 0.7223.

4 Conclusion

In this paper, i presented an empirical evaluation of a number of features and learning algorithms for the task of identifying author profiles. More specifically, the task here is, for a given text, to identify gender, age group and personality traits of its author.

The goal was to validate the use of Information Retrieval-based features to identify personality traits. The results show that they are suitable to the task.

Acknowledgments. I thank to Viviane Pereira Moreira for their help in the final revision of this paper.

References

1. 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)
2. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
3. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling task at PAN 2015. In: Cappellato, L., Ferro, N., Jones, G., SanJuan (eds.) CLEF 2015 Labs and Workshops, Notebook Papers CEUR-WS.org vol. 1391 (2015)
4. Weren, E.R., Kauer, A.U., Mizusaki, L., Moreira, V.P., de Oliveira, J.P.M., Wives, L.K.: Examining multiple features for author profiling. Journal of Information and Data Management 5(3), 266 (2014)
5. Weren, E.R., Moreira, V.P., de Oliveira, J.P.: Exploring information retrieval features for author profiling – notebook for pan at clef 2014. Cappellato et al.[6]
6. Weren, E.R., Moreira, V.P., de Oliveira, J.P.: Using simple content features for the author profiling task. In: Notebook for PAN at Cross-Language Evaluation Forum. Valencia, Spain (2013)