Exploring Information Retrieval features for Author Profiling

Notebook for PAN at CLEF 2014

Edson R. D. Weren, Viviane P. Moreira, and José P. M. de Oliveira

Institute of Informatics UFRGS - Porto Alegre - Brazil
{erdweren,viviane,palazzo}@inf.ufrgs.br

Abstract This paper describes the methods we have employed to solve the author profiling task at PAN-2014. Our goal was to rely mainly on features from Information Retrieval to identify the age group and the gender of the author of a given text. We describe the features, the classification algorithms employed, and how the experiments were run.

1 Introduction

Author profiling deals with the problem of finding as much information as possible about an author, just by analysing a text produced by that author. This is a challenging task which has applications in forensics, marketing, and security [1].

This paper reports on the participation of the INF-UFRGS team at the second edition of the author profiling task, organised in the scope of the PAN Workshop series, which is collocated with CLEF2014. More details about the task and the workshop can be found in [2,5] The task requires that participating teams come up with approaches that take a text as input and predict the gender (male/female) and the age group (18-24, 25-34, 35-49, 50-64, or 64+) of its author.

2 Features

The texts from each author, or *documents*, were represented by a set of 64 features (or attributes), which were divided into five groups. Next, we explain each of these groups.

Length These are simple features that calculate the absolute length of the text.

- `Number of Characters`;
- `Number of Words`; and
- `Number of Sentences`.

Information Retrieval This is the group of features that encode our assumption that authors from the same gender or age group tend to use similar terms and that the distribution of these terms would be different across genders and age groups. The process here was the same as in [6]. The complete set of texts is indexed by an Information Retrieval (IR) System. Then, the text that we wish to classify is used as a query and the k most similar texts are retrieved. The ranking is given by the cosine or Okapi metrics as explained below. We employ a total of 30 IR-based features.

– **Cosine**
`female_cosine_sum, male_cosine_sum, female_cosine_count, male_cosine_count, female_cosine_avg, male_cosine_avg, 18-24_cosine_sum, 25-34_cosine_sum, 35-49_cosine_sum, 50-64_cosine_sum, 65-xx_cosine_sum, 18-24_cosine_count, 25-34_cosine_count, 35-49_cosine_count, 50-64_cosine_count, 65-xx_cosine_count, 18-24_cosine_avg, 25-34_cosine_avg, 35-49_cosine_avg, 50-64_cosine_avg, 65-xx_cosine_avg`.
These features are computed as an aggregation function over the top-k results for each age/gender group obtained in response to a query composed by the keywords in the text that we wish to classify. We tested three types of aggregation functions, namely: count, sum, and average. For this featureset, queries and documents were compared using the cosine similarity (Eq. 1). For example, if we retrieve 100 documents in response to a query composed by the keywords in q, and 50 of the retrieved documents were in the 18-24's age group, then the value for `18-24_cosine_avg` is the the average of the 50 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.

$$cosine(c, q) = \frac{\overrightarrow{c} \cdot \overrightarrow{q}}{|\overrightarrow{c}||\overrightarrow{q}|} \tag{1}$$

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 BM25**
`female_okapi_sum, male_okapi_sum, female_okapi_count, male_okapi_count, female_okapi_avg, male_okapi_avg, 18-24_okapi_sum, 25-34_okapi_sum, 35-49_okapi_sum, 50-64_okapi_sum, 65-xx_okapi_sum, 18-24_okapi_count, 25-34_okapi_count, 35-49_okapi_count, 50-64_okapi_count, 65-xx_okapi_count, 18-24_okapi_avg, 25-34_okapi_avg, 35-49_okapi_avg, 50-64_okapi_avg, 65-xx_okapi_avg .`
Similar to the previous featureset, these features compute an aggregation function (average, sum, and count) over the the retrieved results from each gender/age 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 our experiments, we used $k_1 = 1.2$ and $b = 0.75$.

Readability Readability tests indicate the comprehension difficulty of a text.

– **Flesch-Kincaid readability tests**
We employ two tests that indicate the comprehension difficulty of a text: Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) [4]. They are given by Eqs. 3 and 4. Higher FRE scores indicate a material that is easier to read. For example, a text with a FRE scores between 90 and 100 could be easily read by a 11 year old, while texts with scores below 30 would be best understood by undergraduates. FKGL scores indicate a grade level. A FKGL of 7, indicates that the text is understandable by a 7th grade student. Thus, the higher the FKGL score, the higher the number of years in education required to understand the text. The idea of using these scores is to help distinguish the age of the author. Younger authors are expected to use shorter words and thus have a smaller FKGL and a high FRE.

$$FRE = 206.835 - 1.015\left(\frac{\#words}{\#sentences}\right) - 84.6\left(\frac{\#syllables}{\#words}\right) \quad (3)$$

$$FKGL = 0.39\left(\frac{\#words}{\#sentences}\right) + 11.8\left(\frac{\#syllables}{\#words}\right) - 15.59 \quad (4)$$

Correctness This group of features aims at capturing the correctness of the text.

– Words in the dictionary: ratio between the words from the text found in the OpenOffice US dictionary[1] and the total number of words in the text.
– Cleanliness: ratio between the number of characters in the preprocessed text and the number of characters in the raw text. The idea is to assess how "clean" the original text is.
– Repeated Vowels: in some cases, authors use words with repeated vowels for emphasis. e.g. "I am soo tired". This group of features counts the numbers of repeated vowels (a, e, i, o, and u) in sequence within a word.
– Repeated Punctuation: this features compute the number of repeated punctuation marks (i.e., commas, semi-colons, full stops, question marks, and exclamation marks) in sequence in the text.

Style

– **HTML tags**: this feature consists in counting the number of HTML tags that indicate line breaks
, images <img>, and links <href>.
– **Diversity**: this feature is calculated as the ratio between the distinct words in the text and the total number of words in the text.

[1] http://extensions.openoffice.org/en/project/english-dictionaries-apache-openoffice

Table 1. Top 5 features in terms of Information Gain

Corpus	Lang	Age			Gender		
		Top 5 features	IG	Type	Top 5 features	IG	Type
Twitter	EN	18-24_okapi_sum	0.083	IR	male_okapi_avg	0.160	IR
		50-64_cosine_sum	0.083	IR	25-34_okapi_avg	0.154	IR
		25-34_okapi_sum	0.081	IR	male_okapi_sum	0.153	IR
		25-34_cosine_sum	0.077	IR	35-49_okapi_avg	0.152	IR
		18-24_cosine_sum	0.075	IR	female_okapi_avg	0.140	IR
Twitter	ES	<href>	0.140	Style	number of words	0.183	Length
		25-34_okapi_count	0.136	IR	words in the dictionary	0.157	Correctness
		25-34_cosine_sum	0.129	IR	male_okapi_sum	0.155	IR
		25-34_cosine_count	0.123	IR	diversity	0.149	Style
		50-64_cosine_sum	0.114	IR	male_cosine_sum	0.148	IR
Blog	EN	diversity	0.000	Style	female_cosine_sum	0.156	IR
		male_okapi_sum	0.000	IR	male_okapi_count	0.146	IR
		male_okapi_count	0.000	IR	female_okapi_count	0.137	IR
		female_okapi_count	0.000	IR	female_cosine_count	0.118	IR
		female_okapi_sum	0.000	IR	cleanliness	0.114	Correctness
Blog	ES	25-34_cosine_sum	0.260	IR	number of words	0.251	Length
		words in the dictionary	0.231	Correctness	words in the dictionary	0.226	Correctness
		50-64_okapi_avg	0.224	IR	repeated_e	0.206	Correctness
		50-64_okapi_sum	0.224	IR	50-64_okapi_avg	0.200	IR
		25-34_cosine_count	0.223	IR	male_okapi_sum	0.194	IR
SocialMedia	EN	50-64_cosine_sum	0.122	IR	female_cosine_count	0.008	IR
		50-64_cosine_count	0.122	IR	female_cosine_sum	0.007	IR
		35-49_cosine_count	0.117	IR	female_okapi_count	0.007	IR
		18-24_cosine_count	0.116	IR	male_okapi_count	0.007	IR
		35-49_cosine_sum	0.114	IR	male_cosine_count	0.006	IR
SocialMedia	ES	18-24_okapi_count	0.200	IR	female_cosine_count	0.081	IR
		50-64_okapi_count	0.200	IR	female_cosine_sum	0.079	IR
		18-24_cosine_count	0.193	IR	male_cosine_count	0.071	IR
		35-49_cosine_count	0.191	IR	25-34_cosine_avg	0.053	IR
		18-24_cosine_sum	0.189	IR	female_okapi_count	0.052	IR
Reviews	EN	65-XX_cosine_sum	0.098	IR	female_okapi_count	0.106	IR
		25-34_okapi_count	0.098	IR	male_okapi_count	0.106	IR
		25-34_cosine_count	0.087	IR	female_cosine_count	0.079	IR
		65-XX_cosine_count	0.083	IR	male_cosine_count	0.079	IR
		65-XX_okapi_count	0.082	IR	female_cosine_sum	0.072	IR

3 Usefulness of the Features

In order to evaluate how discriminant each of the 64 features described in Section 2 is, we calculated their information gain with respect to the class. The five highest ranking features for each corpus and each class are shown in Table 1. The vast majority of the most discriminative features is from the IR group. Style, length, and correctness also appear, but at a much lower frequency. For Age-Blogs-EN, none of our features had a good score for information gain. Interestingly, we got the best scores for this corpus on the test data, compared to other groups.

Information gain evaluates each feature independently from each other. However, when selecting the best group of features, we wish to avoid redundant features by keeping features that have at the same time a high correlation with the class and a low intercorrelation. With this aim, we used Weka's [3] subset evaluators to select good subsets of features. These subsets are shown in Table 2. The number of attributes in these

Table 2. Best subset of features for each corpus

Corpus	Lang	Age	Gender
Twitter	EN	18-24_cosine_sum 18-24_cosine_count male_okapi_count 35-49_okapi_count repeated_e repeated_exclamation	male_okapi_sum
Twitter	ES	50-64_cosine_sum 65-XX_cosine_count 25-34_okapi_sum 25-34_okapi_count <href> words_in_dictionary number_of_characters repeated_e repeated_semicolon	male_cosine_sum male_cosine_count words_in_dictionary repeated_exclamation
Blog	EN	male_cosine_avg 50-64_okapi_count <img> repeated_exclamation repeated_interrogation	female_cosine_sum male_cosine_count female_okapi_count
Blog	ES	65-XX_cosine_count 65-XX_cosine_avg 25-34_okapi_sum	repeated_e repeated_exclamation
SocialMedia	EN	female_cosine_avg male_cosine_avg 25-34_cosine_avg 35-49_cosine_avg 18-24_okapi_count 65-XX_okapi_avg FKGL repeated_i repeated_fullstop	male_cosine_count 18-24_cosine_sum 35-49_cosine_count female_okapi_count FRE <img> repeated_exclamation repeated_interrogation
SocialMedia	ES	50-64_cosine_sum 18-24_cosine_count female_okapi_sum male_okapi_count 18-24_okapi_sum 18-24_okapi_count 18-24_okapi_avg <img> number_of_characters repeated_a repeated_ponto	female_cosine_sum male_cosine_avg male_okapi_count 18-24_okapi_count FKGL repeated_a repeated_i repeated_u repeated_exclamation
Reviews	EN	female_cosine_avg 18-24_cosine_sum 65-XX_cosine_sum 65-XX_cosine_count 65-XX_okapi_sum 25-34_okapi_count 65-XX_okapi_count FKGL number_of_characters repeated_i repeated_o repeated_comma repeated_semicolon repeated_exclamation cleanliness diversity	female_cosine_sum 50-64_okapi_count 65-XX_okapi_count diversity repeated_semicolon

subsets varied a lot, from one (Gender-Twitter-EN) to 16 (Age-Reviews-EN). Again, we observed that most features in the subsets are IR-based. Surprisingly, readability features (namely `FKGL`) appear in only two subsets for Age. Style and correctness attributes also appear in the chosen subsets. Also, we noticed that some features that were intended for age, have been selected as useful for gender and vice-versa.

4 Official Experiments

We treated gender and age classification separately. Thus, the features described in the previous section were used to train one classifier for each corpus for gender and age resulting in 14 classifiers. We used Weka [3] to build the machine learning models. A number of algorithms was tested, namely: BayesNet, Logistic, MultilayerPerceptron, SimpleLogistic, LogitBoost, RotationForest, and MetaMultiClass. We chose the algorithm which got the best result for the training data using 10-fold cross-validation. To make such choice, we analysed the results of the classifiers in two scenarios: using all 64 attributes and using just the attributes in the best subset.

The preprocessing consisted basically in tokenisation, removal of tags, and escape characters. No stemming or stopword removal was performed. All training instances were used to generate the model. No attempt to remove noise was taken.

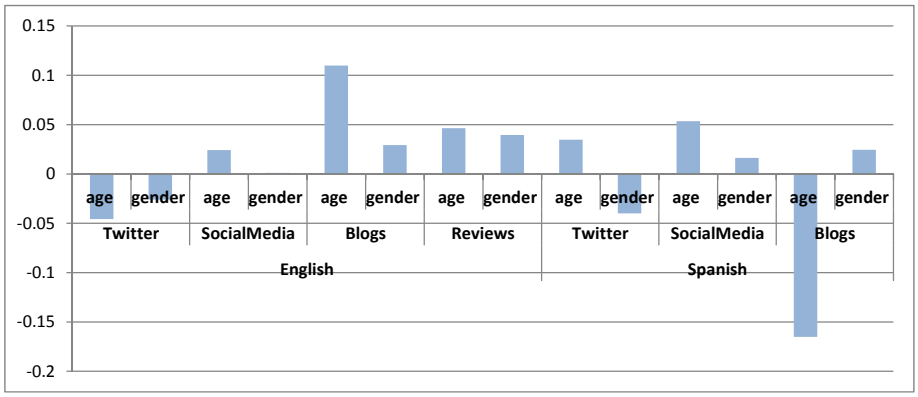
Table 3 shows our official results for both training and test corpora in terms of accuracy. It also shows which classification algorithm was used and whether all attributes or just a subset were used. Most classifiers (11 out of 14) used just the subset of attributes, as their results on the training data outperformed (or got very close to) the results using all attributes.

As expected, results on the training corpora were superior to the results on the test corpora. The biggest drop was for Age-Blog-ES as in this corpus, in which accuracy dropped by half. Interestingly, the results for three corpora were better on the test data (Age-Twitter-ES, Age-Blogs-EN, and Gender-Twitter-ES). We still need to investigate these differences further.

Table 3. Official Results

Age					
Corpus	**Lang**	**Training**	**Test**	**Classifier**	**Attributes**
Twitter	EN	0.5261	0.3312	LogitBoost	Subset
Twitter	ES	0.5056	0.5222	RotationForest	Subset
Blog	EN	0.4558	0.4615	MultiClassClassifier	Subset
Blog	ES	0.5455	0.2500	LogitBoost	Subset
SocialMedia	EN	0.4251	0.3489	Logistic	All
SocialMedia	ES	0.4866	0.4382	Logistic	Subset
Reviews	EN	0.3762	0.3343	Logistic	Subset
Gender					
Corpus	**Lang**	**Training**	**Test**	**Classifier**	**Attributes**
Twitter	EN	0.7876	0.5714	Logistic	Subset
Twitter	ES	0.4494	0.5333	Logistic	All
Blog	EN	0.8299	0.6410	MultilayerPerceptron	Subset
Blog	ES	0.7955	0.5357	RotationForest	Subset
SocialMedia	EN	0.5704	0.5361	SimpleLogistic	Subset
SocialMedia	ES	0.7020	0.6307	SimpleLogistic	All
Reviews	EN	0.7103	0.6778	SimpleLogistic	Subset

Figure 1. Comparison against the mean results of all participants



We also analysed our results compared against the mean of all participants. These are shown in Figure 1. For 9 out of 14 cases, our results were above the mean. The case with the biggest gain was Age-Blogs-EN, in which the advantage was of 31%. In 5 runs, our results were at or below the mean. Our worst scores were for Age-Blogs-ES, in which our loss was of nearly 66%. Adding up all gains and losses, we get a positive result of 10% in relation to the average.

5 Conclusion

This paper describes our participation in the Author Profiling task run in PAN 2014. We used the training data to build classifiers using several machine learning algorithms. Our focus was on exploring Information Retrieval-based features. The official results show that our scores were above the mean for all participants in most cases (9 times out of 14).

Author profiling is a challenging task. Consequently, there are many possibilities for future work. As a first step, once the test data is released, we will further investigate the cases in which our system fails or succeeds in the classification. The goal is to try and establish patterns. We are also interested in testing methods for instance selection to improve our classification models. In addition, we have treated gender and age classification separately as independent problems. However, since some attributes meant to discriminate gender were found useful for age (and vice-versa), we wish to explore the influence of both types of classification into each other.

Acknowledgements: This work has been partially supported by CNPq-Brazil (478979/2012-6). We thank Anderson Kauer for his help in revising this paper. We thank Martin Potthast, Francisco Rangel, and other members of the PAN organising team for their help in getting our software to run.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (Feb 2009)
2. Gollub, T., Potthast, M., Beyer, A., Busse, M., Pardo, F.M.R., Rosso, P., Stamatatos, E., Stein, B.: Recent trends in digital text forensics and its evaluation - plagiarism detection, author identification, and author profiling. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF. Lecture Notes in Computer Science, vol. 8138, pp. 282–302. Springer (2013)
3. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009)
4. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Tech. rep., National Technical Information Service, Springfield, Virginia (Feb 1975)
5. 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)
6. Weren, E.R.D., Kauer, A.U., Mizusaki, L., Moreira, V.P., Oliveira, J.P.M.D., Wives, L.: Examining multiple features for author profiling. Journal of Information and Data Management (JIDM) 5(1) (October 2014), to appear.