Using Simple Content Features for the Author Profiling Task
Notebook for PAN at CLEF 2013

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-2013. Our goal was to use simple features to identify the age group and the gender of the author of a given text. We introduce the features, detail how the classifiers were trained, 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 the author. It has a growing importance in applications such as forensics, marketing and security [1].

This paper reports on the participation of the INF-UFRGS team at the author profiling task which has run for the first time at CLEF2013. In short, the task requires that participating teams come up with approaches that take a given text as input and identify the gender (male/female) and the age group (10s, 20s, 30s) of its author.

As our first attempt in solving the author profiling task, our aim was to design a simple approach in which we exploit features extracted from the contents of the texts. The idea was to try to identify discriminative features and use them in a classifier which predicts the gender and the age group of the author.

2 Identifying Author Profiles

Our underlying assumption was 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. To implement this notion, all conversations were indexed using an Information Retrieval engine and then we treat the conversation we wish to classify as a query. The idea is that the conversations that will be retrieved (*i.e.*, the most similar to the query) will be the ones from the same gender and age group.

The training dataset was composed of conversations (xml files) about various topics grouped by author. Conversations were in English and Spanish and were annotated with the gender and the age group of the author. For a complete description of the dataset, please refer to [5]. Each conversation was represented by a set of features, namely:

−**FeatureSet 1: Cosine**

`Cosine_10s, Cosine_20s, Cosine_30s, Cosine_female, Cosine_male`.
Number of times a conversation from each gender/age group appeared in the top-k ranks for the query composed by the keywords in the conversation. For this featureset, queries

and conversations were compared using the cosine similarity (Eq. 1). For example, if we retrieve 10 conversations in response to a query composed by the keywords in conversation q, and 5 of the retrieved conversations were in the 10's age group, then the value for Cosine_10s is 5.

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

where $\overrightarrow{c}$ and $\overrightarrow{q}$ are the vectors for the conversations 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 conversation c, and $IDF_i = \log \frac{N}{n(i)}$ where N is the total number of conversations in the collection, and $n(i)$ is the number of conversations containing i.

$-$**FeatureSet 2: Okapi**

`Okapi_10s, Okapi_20s, Okapi_30s, Okapi_female, Okapi_male`

Similar to the previous featureset, this is the number of times a conversation from each gender/age group appeared in the top-k ranks for the query composed by the keywords in the conversation. For this featureset, queries and conversations were compared using the Okapi BM25 score (Eq. 2).

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

where $tf_{i,c}$ and IDF_i are as in Eq. 1 $|d|$ is the length (in words) of conversation c, $avgdl$ is the average conversation 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 conversations. In our experiments, we used $k_1 = 1.2$ and $b = 0.75$.

$-$**FeatureSet 3: Flesch-Kincaid readability tests**

There are 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`.

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

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

Training the Classifiers: Four classifiers are necessary, since there are two languages and two dimensions in each (age and gender). We employed a decision-tree classifier. In all cases, the attributes were selected using the BestFirst method.

−**Gender/Spanish**
Training was done on 3K randomly selected conversations.The attributes used were `Cosine_female, Okapi_female`, and `Okapi_male`.
−**Age/Spanish**
Since the number of conversations for the 10's age group was much smaller than the number for the other two classes and classifiers are known to perform better when the number of instances in each class are balanced, we used a method known as *random oversampling*. The method basically selects and replicates random instances from the minority class. According to [2], this approach performs as well as more sophisticated heuristic methods. The attributes used were `Okapi_10s, Okapi_30s, FRE`, and `FKGL`.
−**Gender/English**
Analysing our attributes, we noticed that none of them were good discriminator for gender in English texts. The attributes used were `Cosine_female, Cosine_male, Okapi_female, Okapi_male, FRE`, and `FKGL`.
−**Age/English**
The attributes used were the same as for Spanish. Since the 10s class had fewer instances, random oversampling was applied.

3 Experiments

The steps taken to process the datasets and run our experiments were the following:
1) Pre-process the conversations in the training data to tokenise and remove tags (no stemming or stopword removal was performed).
2) Randomly choose 10% of the conversations to be used as queries.
3) Index the remaining 90% of the pre-processed conversations with a retrieval engine. The system we used was Zettair[1], which is a compact and fast search engine developed by RMIT University (Australia). It performs a series of IR tasks such as indexing and matching. Zettair implements several methods for ranking documents in response to queries and has calculates cosine and Okapi BM25.
4) Compute FeatureSets 1 and 2 using the results from the queries submitted to Zettair. The top 10 scoring conversations were retrieved.
5) Calculate `FRE` and $FKGL$ for the conversations used as queries. The code available from[2] was used.
6) Train the classifiers and generate the decision tree model. Weka [3], was used to build the classification models. It implements several decision tree classification algorithms, we chose J48.
7) Use the trained classifiers to predict the classes of the conversations used as queries.

Once the classifiers are trained, than we can use them to predict the classes for new conversations for which we do not know the age and the gender of the authors. Thus, the conversations from the test data were treated as queries and went through steps 1, 4, 5, and 7.

Table 1 shows our results on the training data. Our best scores were for gender in Spanish (90% correct classification), while our worst results were for gender in En-

[1] `http://www.seg.rmit.edu.au/zettair/`
[2] `http://tikalon.com/blog/2012/readability.c`

Table 1. Results for the training dataset (10-fold cross-validation) and for the test dataset

	Gender/ES	Age/ES	Gender/EN	Age/EN
Correctly Classified	0.91	0.77	0.51	0.55
Precision	0.92	0.76	0.52	0.54
F-measure	0.90	0.77	0.45	0.53
Accuracy - Test Data	0.53	0.46	0.50	0.51

glish (51% correct classification). We attribute this to the fact that in Spanish, most adjectives need to agree with the gender of the author. Thus a woman would say that she is "*cansada*" while a man would say that he is "*cansado*". In English, both would say "*tired*". For age, we also scored better in Spanish. When we look at the test data, however, the scores for Spanish decrease significantly. The most noticeable reduction was for gender, for which only 53% of the conversations were accurately classified. The scores for English remained similar across training and test data. We speculate that this happened because fewer instances were used to generate the Spanish classification models, and they may not have been comprehensive enough to account for all aspects in the data.

4 Conclusion

This paper described our experiments for the author profiling task at PAN-2013. We employed four classifiers which exploit simple features to identify the age group and the gender of authors. This was a preliminary investigation and we plan to continue searching for improvements. Analysing the results from all participating groups, we see that there is still a lot of room for improvement. As future work, we will investigate the use of other features.

Acknowledgements: This work has been partially supported by CNPq-Brazil (478979/2012-6). We thank Martin Potthast and 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. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6(1), 20–29 (Jun 2004)
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. (Feb 1975)
5. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Stamatatos, E., Rosso, P., Stein, B.: Overview of the 5th international competition on plagiarism detection. In: CLEF 2013 Evaluation Labs and Workshops - Working Notes Papers (Sept 2013)