UniNE at CLEF 2016: Author Profiling
Notebook for PAN at CLEF 2016

Mirco Kocher, Jacques Savoy
University of Neuchâtel
rue Emile Argand 11
2000 Neuchâtel, Switzerland
{Mirco.Kocher, Jacques.Savoy}@unine.ch

Abstract. This paper describes and evaluates an author profiling model called SPATIUM-L1. The suggested strategy can be adapted without any problem to different Indo-European languages (such as Dutch, English, and Spanish). As features, we suggest using the *m* most frequent terms of the query text (isolated words and punctuation symbols with *m* at most 200). Applying a simple distance measure and looking at the five nearest neighbors, we can determine the gender (with the nominal values "male" or "female") and the age group (with the ordinal measurement 18-24 | 25-34 | 35-49 | 50-64 | >65). While the labeled data is available for Twitter tweets, the evaluations are based on three test collections from an unknown different genre (blogs, reviews, social media, …) (PAN AUTHOR PROFILING task at CLEF 2016).

1 Introduction

Social network applications produce a big amount of information (*e.g.*, texts, pictures, videos, links) at an unprecedented scale. Texts shared on such sites like Facebook and Twitter have their own characteristics and are difficult to compare with essays, literary texts, or newspaper articles. This is because anybody can publish unrevised content and the compulsion of having a fast interaction. We can observe a large variability related to spelling or grammar. Moreover, new terms tend to appear and emoticons are used frequently to denote the author's emotions or state of mind.

The central question is, if we can detect writings by men and by women from those sources, or if there are no significant differences in their writing style. Similarly, can we detect the features that best discriminate different writings by different age groups? There are some other interesting problems emerging from blogs and social networks such as detecting plagiarism, recognizing stolen identities, or rectifying wrong information about the writer. Therefore, proposing an effective algorithm to the profiling problem presents an indisputable interest.

These author profiling questions can be transformed to authorship attribution questions with a closed set of possible answers. Determining the gender of an author can be seen as attributing the text in question to either the male authors or female authors. Similarly, the age group detection takes one of five groups to attribute the unknown text.

This paper is organized as follows. The next section presents the test collections and the evaluation methodology used in the experiments. The third section explains our proposed algorithm called SPATIUM-L1. In the last section, we evaluate the proposed scheme and compare it to the best performing schemes using three different test collections. A conclusion draws the main findings of this study.

2 Test Collections and Evaluation Methodology

The experiments supporting previous studies were usually limited to custom corpora. To evaluate the effectiveness of different profiling algorithms, the number of tests must be large and run on a common test set. To create such benchmarks, and to promote studies in this domain, the PAN CLEF evaluation campaign was launched (Rangel *et al.*, 2016). Multiple research groups with different backgrounds from around the world have participated in the PAN CLEF 2016 campaign. Each team has proposed a profiling strategy that has been evaluated using the same methodology. The evaluation was performed using the *TIRA* platform, which is an automated tool for deployment and evaluation of the software (Gollub *et al.*, 2012). The data access is restricted such that during a software run the system is encapsulated and thus ensuring that there is no data leakage back to the task participants (Potthast *et al.*, 2014). This evaluation procedure also offers a fair evaluation of the time needed to produce an answer.

During the PAN CLEF 2016 evaluation campaign, three test collections were built. In this context, a problem is simply defined as:

Predict an author's age and gender cross-genre.

In each collection, all the texts matched the same language. The first benchmark is composed of a Dutch collection with the goal to predict the gender. The second is an English corpus and the third is written in Spanish. For the last two, the additional task is to determine the age group. The training data was collected from Twitter. The test data can be blogs, reviews, social media, or any other genre with the exception of Twitter tweets. It was later revealed (after the completion of the task) that the Dutch test corpus contains reviews and both the English and Spanish corpora contain blogs.

Table 1. PAN CLEF 2016 corpora statistics

		Training			Test
Language	Type	Blogs	No. of samples	Mean words	Genre
Dutch	Gender	Twitter	384	2,585	Reviews
English	Gender & Age	Twitter	436	8,120	Blogs
Spanish	Gender & Age	Twitter	250	11,264	Blogs

An overview of these collections is depicted in Table 1. The number of samples from the training set is given under the label "No. of samples" and the mean number of words per sample is indicated under the label "Mean words". A similar test set will then be used in order to be able to compare our results with those of the PAN CLEF

2016 campaign. That datasets remained undisclosed due to the *TIRA* system so we don't have certain information about its size.

When inspecting the Dutch training collection, the mean number of words per question is rather small. Therefore, we can expect the performance to be lower than that for the other two languages. For the Spanish corpus, Table 1 indicates that we have the longest samples to learn the profile from the stylistic features of the author. However, the personal pronouns are not always explicitly specified in this language, (*e.g.*, $we\ can \rightarrow podemos$) and therefore one effective feature able to discriminate the two genders (Pennebaker, 2011) is not fully available (without an effective POS tagger). A relatively higher performance can be assumed in this benchmark. A similar conclusion can be expected with the English collection consisting of the most samples.

When considering the three benchmarks as a whole, we have 1,070 profiles to train our system. When inspecting the distribution of the answers, we can find the same number (535 in training) as male and female profiles. In each of the individual test collections, we can also find a balanced number of male and female profiles. This is not the case for the age group. The two oldest of the five age groups represents only 20% of the English corpus and 17% of the Spanish collection while there are 42% and 50% of the 35-49 year olds as well as 32% and 26% of the 25-34 year olds respectively. This normal distribution is reasonable because only few people (19% as of April 2015[1]) of age 50 or older are using Twitter.

During the PAN CLEF 2016 campaign, a system must provide the answer for each problem in an XML structure. The response for the gender is a fixed binary choice and for the age group one of five fixed entries is expected.

The performance measure is the joint accuracy of the gender and age. This is the number of problems where both the gender and age are correctly predicted for the same problem divided by the number of problems in this corpus. In case no age prediction is requested the joint accuracy is the same as the accuracy of the gender prediction alone.

3 Simple Profiling Algorithm

To solve the profiling problem, we suggest a supervised approach based on a simple feature extraction and distance measure called SPATIUM-L1 (Latin word meaning distance). The selected stylistic features correspond to the top m most frequent terms (isolated words without stemming but with the punctuation symbols). For determining the value of m, previous studies have shown that a value between 200 and 300 tends to provide the best performance (Burrows, 2002; Savoy, 2015). Some profiles were rather short and we further excluded the words only appearing once in the text. This filtering decision was taken to prevent overfitting to single occurrences. The Twitter tweets contained a lot of different hashtags (keyword preceded by a number sign) und numerous unique hyperlinks. To minimize the number of terms with a single occurrence we conflated all hashtags to a single features and combined the morphological variants of Twitter links to another feature. The effective number of

[1] http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/

terms m was set to at most 200 terms but was in most cases well below. With this reduced number the justification of the decision will be simpler to understand because it will be based on words instead of letters, bigrams of letters, or combinations of several representation schemes or distance measures.

In the current study, a profiling problem is defined as a query text, denoted Q, containing blog entries, reviews, or any textual data except Twitter tweets. We then have multiple authors A with a known profile from Twitter tweets. To measure the distance between Q and A, SPATIUM-L1 uses the L1-norm as follows:

$$\Delta(Q, A) = \sum_{i=1}^{m} \left| P_Q[t_i] - P_A[t_i] \right| \quad (1)$$

where m indicates the number of terms (words or punctuation symbols), and $P_Q[t_i]$ and $P_A[t_i]$ represent the estimated occurrence probability of the term t_i in the query text Q or in the author profile A respectively. To estimate these probabilities, we divide the term occurrence frequency (denoted tf_i) by the length in tokens of the corresponding text (n), $\text{Prob}[t_i] = tf_i / n$. Due to the simple difference underlying Equation 1, we do not apply any smoothing procedure to our probability estimation.

To determine the gender and age of Q we take the five nearest neighbors according to SPATIUM-L1 in the m-dimensional vector space and use majority voting. In case five different age groups are returned, we selected the nearest. Since the vector space is spanned by the terms in Q the number of dimensions as well as the bases themselves are likely different from any query text to another and all distances have to be recalculated. This feature selection also means that $\Delta(A, B)$ is not the same as $\Delta(B, A)$ for two profiles A and B. Nevertheless because of the reduced number of features there won't be a performance problem.

4 Evaluation

Our system is based on a supervised approach and we were able to partly evaluate it using older datasets from the PAN CLEF campaign. We took the PAN 2016 corpus (which we know contains Twitter tweets) with the labeled data and validated the English and Spanish performance on various corpora from PAN 2014 while validating the Dutch performance on PAN 2015. In Table 2, we have reported the same performance measure applied during the PAN 2016 campaign, namely the joint accuracy of the gender and age. The expected performance of a random choice would be 50% (or 0.5) for the gender, 20% (or 0.2) for the age, and 10% (or 0.1) for the joint value. The number of problems in those validation corpora can be seen in the column labeled "Size".

Table 2. Evaluation for the *validation* collections

Language	Genre	Size	**joint**	Gender	Age	Runtime (h:m:s)
Dutch	Twitter	34	**0.5588**	0.5588	-	00:00:14
English	Blogs	147	**0.2449**	0.5374	0.4150	00:02:08
English	Review	4,160	**0.1243**	0.4930	0.2478	00:54:02
English	Social media	7,746	**0.1405**	0.5041	0.2755	01:54:56
English	Twitter	306	**0.5297**	0.7647	0.6699	00:04:26
Spanish	Blogs	88	**0.2045**	0.5568	0.3409	00:00:41
Spanish	Social media	1,272	**0.1745**	0.5055	0.3420	00:09:33
Spanish	Twitter	178	**0.4663**	0.7022	0.6629	00:01:31

The algorithm clearly returns the best results for the three Twitter collections because they have the same genre as the labeled corpora. Predicting the gender in the same genre (Twitter) was possible with an accuracy of 76% in English and 70% in Spanish. On the other hand, detecting the true gender cross-genre was achieved with 49% - 56%, which is not a real improvement over random guessing (50%). Therefore, the performance loss when determining the gender is over 33% for English and almost 25% for Spanish. While an arbitrary choice would only get 20% right, the cross-genre age determination is more reliable with up to 42% of the problems correctly classified. But compared to the same genre age prediction the loss of accuracy is around 50%. The text genre has a real impact on the effectiveness and the training set must reflect closely the test set. Due to its large size, we expect the results on the social media and review corpora to be more robust than the ones from the blogs and tweets.

When analyzing the difference between the two genders or the five age groups we can obtain a better understanding of the proposed assignments. From the English training corpus, we learn that female authors use *more pronouns* (especially the second person plural pronouns) and *more hashtags*. The male writers use *more determiners* and have a higher fraction of *complex words* (words having more than 6 characters). Young authors have a heavy usage of the *first person singular pronouns* and "." from the punctuation symbols (full stop; meaning they use rather short sentences and/or add many ellipses indicating intentional omission of words). With the stepwise growing age groups we can observe that the frequency of those features decreases continuously. On the other hand, the *first person plural pronouns* are mostly missing, there are only few *complex words*, and *hyperlinks* are the least frequent in this age group. These sets of features show a constant increase in frequency with higher age groups.

The test set is then used to rank the performance of all 22 participants in the competition. Based on the same evaluation methodology, we achieve the results depicted in Table 3 corresponding to all problems present in the three test collections. As we can see the joint scores on the test corpus are very similar to the cross-genre results from the validation set. For the English and Spanish corpora, we can see a close resemblance to the corresponding results in the validation collections containing blogs. The system seems to perform stable independent of the underlying text collection.

Table 3. Evaluation for the three *testing* collections

Language	**joint**	Gender	Age	Runtime (h:m:s)	Rank
Dutch	**0.5040**	0.5040	-	00:02:27	14
English	**0.2564**	0.5769	0.4103	00:01:18	13
Spanish	**0.1964**	0.5357	0.3393	00:00:30	16

The goal of this year's PAN author profiling task was to determine the age and gender cross-genre. It was still allowed to train the system on other data and to evaluate the performance. We therefore run an experiment for the English and Spanish corpora when using the PAN 2014 blogs as the labeled datasets. This gave a 4% improvement in the gender detection in the English language (62%) and 5% higher accuracy for the age determination in the Spanish corpus (39%). We were free to choose which results should be used in the ranking. In order to ensure the right cross-genre evaluation we selected the results achieved with the provided Twitter data from the current year as it was encouraged by the organizers, even though the performance was slightly lower.

To put those values in perspective we can see in Table 4 our results in comparison with the other 21 participants. The average gender score is the mean over all three languages. But the average age and average joint score is the mean only in the English and Spanish collection as no age prediction was tested in Dutch. The final overall value for the ranking is the mean of those three average values. For the runtime the sum of the runtimes in all three corpora is used. There is also a random (empirical) baseline provided by the organizers. Overall, we are better than the baseline and we are ranked 14 out of 23 approaches.

Table 4. Evaluation over all three *test* collections.

Rank	Run	**Overall**	Average joint	Average Gender	Average Age	Runtime (h:m:s)
1	nissim16	**0.5258**	0.4066	0.6171	0.5538	1:02:23
2	modaresi16b	**0.5247**	0.4066	0.6523	0.5154	0:21:53
3	bilan16	**0.4834**	0.3542	0.6395	0.4565	0:10:50
4	modaresi16a	**0.4602**	0.3121	0.6210	0.4476	0:00:48
5	markov16	**0.4593**	0.3350	0.5954	0.4476	0:08:29
6	bougiatiotis16	**0.4519**	0.3237	0.5956	0.4364	0:01:21
7	dichiu16	**0.4425**	0.2953	0.5948	0.4373	0:04:09
8	devalkeneer16	**0.4369**	0.3031	0.5422	0.4654	0:00:30
9	waser16	**0.4293**	0.2942	0.5703	0.4233	0:06:25
10	bayot16	**0.4255**	0.2608	0.5952	0.4206	0:06:55
11	gencheva16	**0.4015**	0.2532	0.6048	0.3466	0:08:30
12	deneva16	**0.4014**	0.2365	0.6210	0.3466	0:28:24
13	agrawal16	**0.3971**	0.2390	0.5188	0.4334	0:10:14
14	kocher16	**0.3800**	0.2264	0.5389	0.3748	0:04:15
…	…		…	…	…	…
19	Baseline	**0.2747**	0.1074	0.5314	0.1855	
…	…		…	…	…	…

From all the evaluation results[2] we noticed that in the Dutch corpus the gender detection accuracy was generally low. One reason could be that those texts were too short, giving us a small training corpus. Out of all 23 approaches only three teams got a score that is higher than 55% (only one of them higher than 60%) while all other teams do not provide a substantial improvement over random guessing in this language. On the other hand, in both the English and Spanish corpora, half of the contributors predicted the gender in more than 60% of the problems correctly.

The runtime only shows the actual time spent to classify the test set. On *TIRA* there was the possibility to first train the system using the training set which had no influence on the final runtime. Since our system did not need to train any parameters this is negligible for our approach, but it might have been used by other participants.

5 Conclusion

This paper proposes a simple supervised technique to solve the author profiling problem. Assuming that a person's writing style may reveal his/her demographics we propose to characterize the style by considering the *m* most frequent terms (isolated words and punctuation symbols). This choice was found effective for other related tasks such as authorship attribution (Burrows, 2002). Moreover, compared to various feature selection strategies used in text categorization (Sebastiani, 2002), the most frequent terms tend to select the most discriminative features when applied to stylistic studies (Savoy, 2015). In order to take the profiling decision, we propose using the five nearest neighbors according to a simple distance measure called SPATIUM-L1 based on the L1 norm.

The proposed approach tends to perform well in English across different genres. The performance on the Spanish dataset was acceptable, but in Dutch the gender detection did not provide considerable improvement over the baseline. Those results were expected from the validation corpora. Such a classifier strategy can be described as having a high bias but a low variance (Hastie *et al.*, 2009). Even if the proposed system cannot capture all possible stylistic features (bias), changing the available data does not modify significantly the overall performance (variance).

Moreover, the proposed profiling could be clearly explained because it is based on a reduced set of features on the one hand and, on the other, those features are words or punctuation symbols. Thus the interpretation for the final user is clearer than when working with a huge number of features, when dealing with *n*-grams of letters or when combing several similarity measures. The SPATIUM-L1 decision can be explained by large differences in relative frequencies (or probabilities) of frequent words, usually corresponding to functional terms.

This year's biggest challenge in the PAN author profiling task were clearly the cross-genre datasets. The testing of the proposed systems was performed on writings from a dissimilar genre than the provided labeled texts. Nevertheless, we were able to show that there exists a difference in writing style between the genders and the tested age groups which is not bound to the genre and can be transferred to other documents.

[2] http://www.tira.io/task/author-profiling/

To improve the current classifier, we will investigate the effect of other distance measures as well as other feature selection strategies. In this latter case, we want to maintain a reduced number of terms. In a better feature selection scheme, we can take account of the underlying text genre, as for example, the most frequent use of personal pronouns in narrative texts. As another possible improvement, we can ignore specific topical terms or character names appearing frequently in an author profile, terms that can be selected in the feature set without being useful in discriminating between authors. One might also try to exploit PAN specific properties such as the requirement for equally distributed male/female problems or the probability to find a normal distribution of the age groups.

Acknowledgments

The author wants to thank the task coordinators for their valuable effort to promote test collections in author profiling. This research was supported, in part, by the NSF under Grant #200021_149665/1.

6 References

1. Burrows, J.F. 2002. Delta: A Measure of Stylistic Difference and a Guide to Likely Author-ship. *Literary and Linguistic Computing*, 17(3), 267-287.
2. Gollub, T., Stein, B., & Burrows, T. 2012. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh, B., Callan, J., Maarek, Y., & Sanderson, M. (eds.) SIGIR. *The 35th International ACM*, 1125–1126.
3. Hastie, T., Tibshirani, R., & Friedman, J. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer-Verlag: New York (NY).
4. Pennebaker, J.W. 2011. *The Secret Life of Pronouns. What our Words Say about us.* Bloomsbury Press: New York (NY).
5. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., & Stein, B. 2014. Improving the Reproducibility of PAN's Shared Tasks: - Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Handbury, A., & Toms, E. (eds.) CLEF. *Lecture Notes in Computer Science*, vol. 8685, 268–299. Springer: Heidelberg.
6. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. 2016. Evaluations Concerning Cross-genre Author Profiling. In Working Notes Papers of the CLEF 2016 Evaluation Labs, *CEUR Workshop Proceedings*. CEUR-WS.org.

7. Savoy, J. 2015. Comparative Evaluation of Term Selection Functions for Authorship Attribution. *Digital Scholarship in the Humanities*, 30(2), 246-261.
8. Sebastiani, F. 2002. Machine Learning in Automatic Text Categorization. *ACM Computing Survey*, 34(1), 1-27.