# UniNE at CLEF 2015: Author Profiling
## Notebook for PAN at CLEF 2015

**Mirco Kocher**
University of Neuchâtel
rue Emile Argand 11
2000 Neuchâtel, Switzerland
Mirco.Kocher@unine.ch

**Abstract.** This paper describes and evaluates an effective author profiling model called SPATIUM-L1. The suggested strategy can be adapted without any problem to different languages (such as Dutch, English, Italian, and Spanish) in Twitter tweets. As features, we suggest using the 200 most frequent terms of the query text (isolated words and punctuation symbols). Applying a simple distance measure and looking at the three nearest neighbors, we can determine the gender (with the nominal values male and female), the age group (with the ordinal measurement 18-24|25-34|35-49|>50), and the Big Five personality traits (extraversion, neuroticism, agreeableness, conscientiousness, and openness on an interval scale containing eleven items). Evaluations are based on four test collections (PAN AUTHOR PROFILING task at CLEF 2015).

## 1 Introduction

Do men write like women, or are there significant differences in their writing style? What are the features that best discriminate different writings by different age groups? Is it possible to detect reliably somebody's personality traits based on a text excerpt? With the Internet, the number of anonymous or pseudonymous texts is increasing and in many cases we face a single author. There are some 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 four groups to attribute the unknown text. Uncovering the Big Five personality traits takes this approach even further by selecting for each factor one of eleven groups (from -0.5 to +0.5 with a step size of 0.1).

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 four 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 [6]. Multiple research groups with different backgrounds from around the world have participated in the PAN CLEF 2015 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 [2]. 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 [4].

During the PAN CLEF 2015 evaluation campaign, four test collections were built. In this context, a problem is defined as:

*Predict an author's demographics from her Twitter tweets.*

In each collection, all the texts matched the same language. These four benchmarks are composed of a Dutch and Italian collection with the task to predict the gender and personality traits and an English and Spanish corpus with the additional goal to determine the age group. The data was collected from Twitter by means of advertising campaign. The gender and age group is therefore user specified while the personality trait labels are gold standard self-assessed with the BFI-10 test [5] and then normalized between -0.5 and +0.5. We will assume that this will reveal accurately the personality traits.

| | | Training | | Test |
|---|---|---|---|---|
| Language | Type | No of Samples | Mean words | No of Problems |
| Dutch | Gender & Personality | 34 | 593 | ~32 |
| English | Gender, Age & Personality | 152 | 527 | ~142 |
| Italian | Gender & Personality | 38 | 638 | ~36 |
| Spanish | Gender, Age & Personality | 100 | 665 | ~88 |

**Table 1.** PAN CLEF 2015 corpora statistics

An overview of these collections is depicted in Table 1. The training set will be used to evaluate our approach and the test set will be used in order to be able to compare our results with those of the PAN CLEF 2015 campaign. 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". For the test set we estimated the number of problems from the accuracy scores of all participants (subject to integer number of correct answers and same number of problems). The 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 number of samples available is rather small. Similarly the Italian collection only provides 38 samples. To predict the value of a personality trait we have in mean only three examples. Therefore, we can expect the performance for these languages to be lower than that for the other 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. 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 four benchmarks as a whole, we have 298 problems to solve and 324 to train our system. When inspecting the distribution of the answers, we can find the same number (149 in test and 162 in training) as male or 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 or the personality traits. The highest of the four age groups represents only 8% of the English corpus and 10% of the Spanish collection while there are 39% and 46% of the 24-34 year olds respectively. The positive skew of this distribution is reasonable because only few people (16% as of October 2014[1]) of age 50 or older are using Twitter. The sampling also suffers from under-coverage of the author's personality traits. For instance for the openness factor in the rather large English and Spanish corpora we cannot find any value of -0.2 or lower and therefore missing the four lowest items on the interval scale. The small Dutch collection even misses samples from the first six items for this trait. Furthermore none of the traits in any of the corpora contained the value -0.4 or -0.5.

During the PAN CLEF 2015 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 four fixed entries is expected. The Big Five personality traits are each answered with a value between -0.5 and +0.5.

As performance measure, two evaluation measures were used during the PAN CLEF campaign. The first 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.

As a measure for the personality traits, the PAN CLEF campaign adopts the Root Mean Square Error (RMSE). This evaluation measure takes into account how far off the predicted value is compared to the values actually observed independent of the direction. The exact formulation is given in Equation 1 with a minimal value of 1.0 and 0.0 as an optimum value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{f}_i - f)^2}{n}} \qquad (1)$$

in which $n$ is the number of problems, $f$ the actual correct trait factor value, and $\hat{f}_i$ the predicted value for problem $i$ of this trait factor. This measure differentiates between a value close to the actual value and an answer far away from the truth. The overall RMSE is the arithmetic mean of the RMSE of the five factors in the Big Five personality trait model.

[1] http://jetscram.com/blog/industry-news/social-media-user-statistics-and-age-demographics-2014/

## 3 Simple Profiling Algorithm

To solve the profiling problem, we suggest a supervised approach based on a simple feature extraction and distance metric called SPATIUM-L1 (Latin word meaning distance). The selected stylistic features correspond to the top $k$ most frequent terms (isolated words without stemming but with the punctuation symbols). For determining the value of $k$, previous studies have shown that a value between 200 and 300 tends to provide the best performance [1, 7]. 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 terms $k$ 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 multiple Twitter tweets. We then have multiple authors A with a known profile. To measure the distance between Q and A, SPATIUM-L1 uses the L1-norm as follows:

$$\Delta(Q,A) = \sum_{i=1}^{k} \left| P_Q[t_i] - P_A[t_i] \right| \tag{2}$$

where $k$ 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 ($n$), $\text{Prob}[t_i] = tf_i / n$.

To determine the gender and age of Q we take the three nearest neighbors according to SPATIUM-L1 in the $k$-dimensional vector space and use majority voting. In case three different age groups are returned we selected the nearest. For each of the five personality traits we use the arithmetic mean of the corresponding traits of those same three candidates. 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. 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 evaluate it using a modified leave-one-out method on the training set. Instead of retrieving the three nearest neighbors we returned four candidates, but ignored the closest profile. The nearest sample was in fact the query text with a L1-disance of zero and thus could also serve as a check of correctness. In Table 2, we have reported the same performance measure applied during the PAN CLEF campaign, namely the global score, which is the mean of the joint accuracy and the overall RMSE subtracted to 1.

| Language | **Global** | joint | RMSE | Runtime (h:m:s) |
|---|---|---|---|---|
| Dutch | **0.8116** | 0.7353 | 0.1121 | 00:00:03 |
| English | **0.7415** | 0.6382 | 0.1551 | 00:00:04 |
| Italian | **0.7854** | 0.7105 | 0.1397 | 00:00:01 |
| Spanish | **0.7530** | 0.6500 | 0.1441 | 00:00:02 |

**Table 2.** Evaluation for the four *training* collections

The algorithm returns the best results for the Dutch collection with a global score of 0.8116 closely followed by the Italian corpus. One has to consider that those two datasets did not require a prediction for the age. Therefore the joint accuracy of the English and Spanish corpora is heavily influenced by an additional category in question. This makes a direct comparison between the languages difficult. Furthermore the former two only contained few problems while the latter two predictions are based on a bigger collection and thus we expect it to be more stable in the second case.

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 the 298 problems present in the four test collections. As we can see the global scores on the test corpus is only slightly higher than the results from the training set. The system seems to perform stable independent of the underlying text collection.

| Language | **Global** | joint | RMSE | Runtime (h:m:s) | Rank |
|---|---|---|---|---|---|
| Dutch | **0.8469** | 0.8125 | 0.1186 | 00:00:01 | 6 |
| English | **0.7037** | 0.5563 | 0.1489 | 00:00:04 | 8 |
| Italian | **0.8260** | 0.7778 | 0.1259 | 00:00:01 | 4 |
| Spanish | **0.7735** | 0.6705 | 0.1235 | 00:00:02 | 4 |

**Table 3.** Evaluation for the four *testing* collections

To put those values in perspective we can see in Table 4 our results in comparison with the other 21 participants using macro-averaging. We have also added a baseline from the training collections corresponding to a system that always produces the same answer. The gender is fixed as *female*, the age is set to *25-34* which is the mode of the age groups, and *0.2* is chosen for all personality traits according to the median (and mode) of the training data.

| Rank | Run | **Global** | joint | RMSE | Runtime (h:m:s) |
|---|---|---|---|---|---|
| 1 | alvarezcarmona15 | **0.8404** | 0.7895 | 0.1087 | 00:02:32 |
| 2 | gonzalesgallardo15 | **0.8346** | 0.8001 | 0.1308 | 00:13:45 |
| 3 | grivas15 | **0.8078** | 0.7882 | 0.1727 | 00:04:07 |
| 4 | kocher15 | **0.7875** | 0.7043 | 0.1292 | 00:00:08 |
| … | … | **…** | … | … | … |
| 20 | Baseline<br>(female, 25-34, 0.2) | **0.5934** | 0.3569 | 0.1702 | 00:00:00 |
| … | … | **…** | … | … | … |

**Table 4.** Evaluation over all four test collections using macro-averaging for the effectiveness measures and the sum for the runtimes.

From all the evaluation results[2] we noticed that gender detection in the Spanish corpus was very high with a median accuracy of almost 85%. In this language the grammatical gender of a noun affects the form of determiners, adjectives, and pronouns related to it. Since Twitter tweets are often about the author him/herself the classification of the gender can be simplified. On the other hand the gender recognition in the Dutch collection has a median accuracy of just 65%. Gender in Dutch is more complicated. The formal and written tradition mostly distinguishes masculine and feminine nouns, but in informal speech (and therefore for tweets too) the distinction disappeared and a common gender with the same inflections and pronouns is used.

We also noted that determining the value of the neuroticism factor seems to be the most complicated in all four languages. In mean the other four personality traits are determined with an RMSE of about 0.15, but in this case the RMSE was around 0.2. It could be possible that the tendency to experience negative emotions (such as anger, anxiety, or depression) is more complicated to determine from written text or that people tend to give less reliable answers on self-assessment tests.

Another pertinent observation is the fast runtime of our system in comparison with other solutions. The median execution time of the other systems is over ten minutes. The practical applicability of such systems could be questioned. 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 personality traits we propose to characterize the style by considering the $k$ most frequent terms (isolated words and punctuation symbols). This choice was found effective for other related tasks such as authorship attribution [1]. Moreover, compared to various feature selection strategies used in text categorization [8], the most frequent terms tend to select the most discriminative features when applied to stylistic studies [7]. In order to take the profiling decision, we propose using the three nearest neighbors according to a simple distance metric called SPATIUM-L1 based on the L1 norm.

The proposed approach tends to perform very well in four different languages (Dutch, English, Italian, and Spanish) for Twitter tweets. Such a classifier strategy can be described as having a high bias but a low variance [3]. 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

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

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.

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. As a further alternative we could consider the distance between the *k* nearest neighbors and the query text when determining the personality traits for a weighted mean instead of the arithmetic mean. We might also try to exploit PAN specific properties such as the requirement for equally distributed male/female problems or the probability to find a right skewed distribution of the age groups.

## Acknowledgments

The author wants to thank the task coordinators for their valuable effort to promote test collections in authorship attribution. 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. 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.

5. Rammstedt, B., & John, O.P. 2007. Measuring personality in one minute or less: A 10 item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41, 203-212.

6. Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In Capellato, L., Ferro, N., Gareth, J., & San Juan, E. (Eds.) *CLEF-2015 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

7. Savoy, J. 2015. Comparative Evaluation of Term Selection Functions for Authorship Attribution. *Digital Scholarship in the Humanities, to appear* (dx.doi.org/10.1093/llc/fqt047).

8. Sebastiani, F. 2002. Machine Learning in Automatic Text Categorization. ACM *Computing Survey*, 34(1), 1-27.