UniNE at CLEF 2017: Author Profiling Reasoning

Notebook for PAN at CLEF 2017

Mirco Kocher and Jacques Savoy

Computer Science Dept., University of Neuchâtel, Switzerland
{Mirco.Kocher, Jacques.Savoy}@unine.ch

Abstract. This paper describes and evaluates a supervised author profiling model. The suggested strategy can be adapted without any problem to various languages (such as Arabic, English, Spanish, and Portuguese). 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 nearest text profiles, we can determine the gender (with the nominal values "male" or "female") and the language variety (e.g., in Spanish the nominal values "Argentina", "Chile", "Colombia", "Mexico", "Peru", "Spain", or "Venezuela"). The training and test data is available for Twitter tweets (PAN AUTHOR PROFILING task at CLEF 2017). An analysis of the top ranked terms from a feature selection method allows a better understanding of the proposed assignments and presents typical writing styles for each category.

1 Introduction

Social network applications produce a big amount of information (e.g., texts, pictures, videos, and links) at an unprecedented scale. Texts shared on such sites like Facebook and Twitter have their own characteristics vastly different from 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 and grammar. Moreover, new terms tend to appear and emoji are used frequently to denote the author's emotions or state of mind.

The central question is, if we can detect writings by the author's gender from those sources, and what are the significant differences between man and women in their writing style. Similarly, can we detect the features that best discriminate different writings by different language varieties? The spelling difference between British English and American English is well defined, but can we detect a variation from the US to Canada, or Ireland and Great Britain, and can we discriminate between New Zealand and Australia? Furthermore, since profiling is based on Twitter tweets, the spelling may not always be perfect, and more sociocultural traits could be detected. 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 or female authors. Similarly, the language variety detection takes one of seven groups to attribute an unknown Spanish 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. Then, we evaluate the proposed scheme and compare it to the best performing schemes using four different test collections. In the last section, we explain the decisions taken and extract typical writing styles for each category. 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 2017 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 [5]. This evaluation procedure also offers a fair evaluation of the time needed to produce an answer.

During the PAN CLEF 2017 evaluation campaign, three test collections were built. In this context, a problem is simply defined as:

Predict an author's language variety and gender from tweets.

In each collection, all the texts matched the same language. The first benchmark is composed of an Arabic collection with the goal to predict four language varieties. The second is an English corpus containing six varieties, the third is written in Spanish and covers seven different varieties, while the last collection is in Portuguese based on two language varieties. In all corpora, the additional task is to determine the author's gender. The training data was collected from Twitter. This year, everyone had access to the test data twice. This means we can train and test a basic approach, improve it, and test it again for the second and final run.

An overview of these collections is depicted in Table 1. The number of samples from the training set is given under the label "Samples" (each sample is a set of tweets) and the mean number of tokens (isolated words and punctuation symbols) per sample is indicated under the label "Terms". A similar test set will then be used to be able to compare our results with those of the PAN CLEF 2017 campaign. That datasets remained mostly undisclosed due to the *TIRA* system so we don't have information about the average number of words per sample, but we expect a similar distribution.

When considering the four benchmarks, we have 11,400 profiles in total to train our system. When inspecting the distribution of the answers, we can find the same number (5,700 in training) of female and male profiles. In each of the individual test collections, we can also find a balanced number of female and male profiles. The same is the case for the language varieties, where each group has 600 samples. During the PAN CLEF 2017 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 language variety one of the fixed entries is expected.

Table 1. PAN CLEF 2017 corpora statistics.

Corpus	Language Varieties	Training		Testing
		Samples	Terms	Samples
Arabic	Egypt, Gulf, Levantine, Maghrebi	2,400	1,241.8	1,600
English	Australia, Canada, Great Britain, Ireland, New Zealand, United States	3,600	1,628.5	2,400
Spanish	Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela	4,200	1,472.3	2,800
Portuguese	Brazil, Portugal	1,200	1,202.3	800

The final performance measure is the joint accuracy of the gender and variety. This is the number of problems where both the gender and language variety are correctly predicted for the same problem divided by the number of problems in this corpus.

3 Profiling Algorithm

To solve the profiling problem, we suggest a supervised approach based on a feature extraction and distance measure. The selected stylistic features correspond to the top m best terms (isolated words without stemming but with the punctuation symbols) calculated by the gain ratio formula as shown in Equation 1.

$$GainRatio(a,b,c,d) = \frac{a}{n}\log_2\left(\frac{a*n}{(a+b)*(a+c)}\right) + \frac{c}{n}\log_2\left(\frac{c*n}{(a+c)*(c+d)}\right) \quad (1)$$

where a, b, c, d, and n are used as indicated in Table 2. For instance, a represents the frequency of a given term ω (e.g., "the" or "people") in each class Γ (e.g., "female" or "Mexico") while d is the sum of all other terms in all other classes.

Table 2. Contingency table for a term ω and in a class Γ.

	Γ	¬Γ	
ω	*a*	*b*	*a*+*b*
¬ω	*c*	*d*	*c*+*d*
	a+*c*	*b*+*d*	*n*

For determining the number of useful features denoted *m*, previous studies have shown that a value between 200 and 300 tends to provide the best performance [1, 7]. 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 feature and combined the morphological variants of Twitter links to another feature. The effective number of terms *m* was set to the 100 highest terms for each gender and 70 highest terms for each language variety. In the first run we also included the 10 lowest ranked terms as a counter indication for a given category, while this was omitted in the second run. Since there is some overlap when combining the highest ranked terms of one class with another, the length of the generated feature list was below 400 even for the Spanish collection containing seven different language classes. 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 a set of Twitter tweets. We then have multiple authors *A* with a known profile. To measure the distance between *Q* and *A*, in the first run we used a variant of the L^1-norm called Canberra as shown in Equation 2, while in the second run we used a variant of the L^2 norm called Clark as shown in Equation 3:

$$\Delta_{Canberra}(Q,A) = \sum_{i=1}^{m} \frac{|P_Q[f_i]-P_A[f_i]|}{P_Q[f_i]+P_A[f_i]} \qquad (2)$$

$$\Delta_{Clark}(Q,A) = \sqrt{\sum_{i=1}^{m} \left(\frac{|P_Q[f_i]-P_A[f_i]|}{P_Q[f_i]+P_A[f_i]}\right)^2} \qquad (3)$$

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*), Prob[t_i] = tf_i / *n*. Due to the simple difference underlying the two Equations, we do not apply any smoothing procedure to our probability estimation.

To determine the gender and variety of *Q* we take the *k* nearest neighbors in the *m*-dimensional vector space and use majority voting. In case there is a tie between multiple language varieties, we selected the nearest group among them. In the first run, the parameter *k* was set to *k=9*. In the second run we increased *k* to *k=15* for the two smaller collections (Arabic and Portuguese) and set *k=25* for the two bigger corpora (English and Spanish). This decision was taken because of the relatively large amount of data available, and to gain a more stable system less affected by outliers or the imperfection of Twitter tweets. A summarization of all parameters in the two runs is presented in Table 3.

Table 3. Parameter summarization.

Parameter		First Run	Second Run
Distance		Canberra	Clark
Feature selection method		Gain Ratio	Gain Ratio
m features	each gender	100 highest 10 lowest	100 highest 0 lowest
	each variety	70 highest 10 lowest	70 highest 0 lowest
k neighbors		9 in AR & PT 9 in EN & SP	15 in AR & PT 25 in EN & SP

4 Evaluation

Our system is based on a supervised approach and we could evaluate it using a modified leave-one-out approach on the training set. Instead of retrieving the *k* nearest neighbors, we returned *k+1* candidates, but ignored the closest profile. The nearest sample was in fact the query text with a distance of zero and thus could also serve as a check of correctness. In Table 4a and Table 4b, we have reported the same performance measures applied during the PAN 2017 campaign, namely the joint accuracy of the gender and language variety.

Table 4a. Evaluation for the four *training* collections with the *first* run.

Language	**Joint**	Gender	Variety
Arabic	0.5021	0.6854	0.7175
English	0.3772	0.6928	0.5411
Spanish	0.4117	0.6445	0.6419
Portuguese	0.7600	0.7742	0.9808
Overall	0.5128	0.6992	0.7203

Table 4b. Evaluation for the four *training* collections with the *second* run.

Language	**Joint**	Gender	Variety
Arabic	0.5292	0.6954	0.7375
English	0.4581	0.7192	0.6392
Spanish	0.4762	0.6745	0.7169
Portuguese	0.7850	0.7967	0.9842
Overall	0.5621	0.7215	0.7695

The algorithm clearly returns the best results for the Portuguese collection as a result of both the high gender detection accuracy and the high language variety prediction accuracy. With the leave-one-out approach and with the large size of all collections, we expect the results to be robust and a good prediction for the test dataset.

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 5a and Table 5b corresponding to our two runs for all problems present in the four test collections. As we can see the joint scores on the test corpus are very similar to the training results. For the Arabic and English corpora, we can see a close resemblance to the corresponding results in the training collections. In the Spanish collection, the test performance is marginally higher (+3.5% change, +8.4% difference), while for the Portuguese dataset, the results are slightly lower (-2.8% change, -3.5% difference). Overall, the system seems to perform stable independent of the underlying text collection.

Table 5a. Evaluation for the four *testing* collections with the *first* run.

Language	**Joint**	Gender	Variety
Arabic	0.5119	0.6781	0.7106
English	0.3879	0.6996	0.5596
Spanish	0.4464	0.6711	0.6611
Portuguese	0.7400	0.7625	0.9713
Overall	0.5216	0.7028	0.7257

Table 5b. Evaluation for the four *testing* collection with the *second* run.

Language	**Joint**	Gender	Variety
Arabic	0.5206	0.6913	0.7188
English	0.4650	0.7163	0.6521
Spanish	0.4971	0.6846	0.7211
Portuguese	0.7575	0.7788	0.9725
Overall	0.5601	0.7178	0.7661

This year, there were 22 participants and the task organizers provided 3 additional baselines[1]. To put our achieved performance values from Table 5b in perspective we can see in Table 6 our results in comparison with the best participant, the three baselines, and the mean performance of all participations scores. The columns with the average gender score, the average language variety score, and the average joint score are each the mean over all four languages. The final overall value for the ranking is the mean of those three average values. Overall, we are at rank 16[2],which is above the average PAN scores and two of the provided baselines.

Table 6. Evaluation over all four *test* collections.

Approach	Average Gender	Average Variety	Average Joint	**Overall**
Basile et al.	0.8253	0.9184	0.8361	**0.8599**
LDR baseline	0.7325	0.9187	0.7750	**0.8087**
Kocher & Savoy	0.7178	0.7661	0.6813	**0.7217**
PAN average	0.6561	0.7099	0.6333	**0.6664**
BOW baseline	0.6763	0.6907	0.6195	**0.6622**
STAT baseline	0.5000	0.2649	0.2991	**0.3547**

[1] http://pan.webis.de/clef17/pan17-web/author-profiling.html
[2] http://www.tira.io/task/author-profiling/

5 Decision Explanation

When analyzing the top ranked terms from the feature selection method between the two genders or the language variety groups we can obtain a better understanding of the proposed assignments. The gain ratio selects both features that are overly present in each category as well as features where it's rarity is a counterindication of a given category. Thus, the selected features are usually the same for both gender classes. To present typical features for each category individually, we use the Mutual Information for the terms in Table 7. This feature selection method assigns a high value only to the overused terms, which gives us a clearer differentiation[3].

In many cases, the different usage of geographical and topical terms can explain the decision for the classification. Some location related terms are for instance in Arabic (الكويت = Kuwait, الاردن = Jordan, طرابلس = Tripoli, الجزائر = Algeria, تونس = Tunis, التونسي = Tunisia), in English (*Canberra*, *Sydney*, *aust*, *Adelaide*, *jp*, *aus*, *Vancouver*, *Toronto*, *Edinburgh*, *Glasgow*, *Bristol*, *Dublin*, *Ireland*, *Belfast*, *Wellington*, *Auckland*, *nz*, *Zealand*, *Dunedin*, *DC*), in Spanish (*chilenos*, *Bogotá*, *Cali*, *Medellín*, *mx*, *Monterrey*, *Lima*, *peruano(s)*, *Perú*, *Peru*, *Alcalá*, *Cataluña*, *Zulia*, *Caracas*, *venezolanos*), and in Portuguese (*Brasil*, *Portugal*).

For topical words, we have different examples in Arabic (مدرب = coach; الدوري = league; #صلاة = #Prayer), in English (*NHL*; *makeup*; *Microsoft*), in Spanish (*lagos* = lakes, *forestales* = forests, *incendios* = fires, *viña* = vineyard, *medicinas* = medicines), and in Portuguese (*campeonato* = championship, *jogador* = player, *ranking*).

Additionally, names of famous people in politics, music, and sports appear frequently, such as in Arabic (عايزة = Aiza), in Spanish (*Zidane*, *Macri*, *Piñera*, *Duarte*, *Goya*, *Rajoy*), in English (*Turnbull*, *Abbott*, *Malcom*, *Reuters*, *Jedward*, *Byrne*, *Conor*, *Ethan*), and in Portuguese (*Eduardo*).

Very frequent terms such as pronouns and determiners also appear in the top 10 highest ranked terms. There are examples in Arabic (إنى = I am; انتى = you; ده = this), in Spanish (*nosotras* = we; *vos*, *os*, *vosotros* = you), and in Portuguese (*vc*, *você* = you; *tô* = I am).

Furthermore, the frequent appearance of various heart shaped emoji in the female categories of Table 7 in all four languages confirms previous findings that women tend to use more expressions related to social and emotion words than men [4].

[3] Some terms depend on the context in which they are used and can't be translated accurately.

Table 7. Top 10 terms selected using mutual information

Category		Top terms (space separated)
Arabic	Female	عارفة عايزة سلمى عايزه ماما حبيبتي 💗 عارفه 🎶 إنى
	Male	مدرب الدوري حازم تغريدة مدريد ' شرح video حان liked
	Egypt	دى اللى تانى كام يعنى انتى النهاردة دلوقتي بقت ده
	Gulf	الحين شلون محد الكويت دايم فيني مافيه ' بو كفو
	Levantine	اشي الاردن بدك الأردن هاي هيك حدا هلأ بده منيح
	Maghrebi	ليبيا طرابلس الجزائر هدا تونس #صلاة التونسي ⭐ معجزة تاع
English	Female	leo taurus virgo xxx 💕 😘 ✨ makeup xx bingo
	Male	)' badge arsenal earned league microsoft wire players developer rangers
	Australia	canberra turnbull sydney aust abbott malcolm jp adelaide scarlet aus
	Canada	vancouver toronto canadians canadian 🇨🇦 220 nhl txt canvas rsvp
	GB	edinburgh filthy glasgow factual unlimited reuters mural bristol drafted gems
	Ireland	dublin ireland commented irish scorpio jedward byrne conor capricorn belfast
	NZ	wellington auckland nz kiwi zealand ✌ dunedin earthquake)' roundup
	US	gorsuch emerald dems ethan scotus dc aca obamacare infamous nsc
Spanish	Female	♡ orgullosa cansada pedidos nosotras angie dormida 💗 siiii celosa
	Male	dt jugó rival refuerzos delantero clubes colo cont zidane libertadores
	Argentina	posta hs podes vos orto lpm pelotuda bue pelotudo macri
	Chile	wn piñera colo lagos incendios po metropolitana forestales viña chilenos
	Colombia	bogotá bogota uribe corridas boletas falcao lleras cali plebiscito medellín
	Mexico	neta mx éxico monterrey pinches duarte hidalgo slim pri 🇲🇽
	Peru	ppk lima peruanos soles ptm perú peru oe muni peruano
	Spain	psoe os vosotros enhorabuena goya pp rajoy vuestro alcalá cataluña
	Venezuela	mud zulia vzla caracas chavista an venezolanos medicinas chavismo hampa
Portuguese	Female	sozinha cansada obrigada ranking achavam ❤❤❤ enviadas simpático apaixonada acordada
	Male	link eduardo | obrigado milhões by •): jogador campeonato
	Brazil	tô fazendo vc você kkkkk at kkkkkk brasil querendo assistir
	Portugal	tou portugal isto cenas crlh gira xd merdas percebo lol

6 Conclusion

This paper proposes a supervised technique to solve the author profiling problem. If a person's writing style may reveal his/her demographics we propose to characterize the style by considering terms (isolated words and punctuation symbols) selected using the gain ratio method. To take the profiling decision, we propose using the *k* nearest neighbors according to a distance measure based on the L^1 or L^2 norm.

The proposed approach tends to perform very well in Portuguese Twitter tweets for both gender and language variety prediction. The performance of the gender detection in Arabic, English, and Spanish was acceptable, while the language variety classification was good considering the large number of categories. The final results on the test collections were as expected from the training corpora, indicating that no over-fitting occurred. 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 can 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 decision can be explained either by large differences in relative frequencies (or probabilities) of frequent words (usually corresponding to functional terms), topical words, or geographical terms. We were able to show that there exists a difference in writing style between the genders and the tested language variety groups.

To improve the current classifier, we could investigate the effect of other feature selection strategies. In this case, we want to maintain a reduced number of terms but we can take more account of the underlying text genre, as for example, the frequent use of emoji in tweets contain more implicit expressions and meanings. Furthermore, we could use external resources to harvest geographical names related to the different countries and regions to facilitate the language variety prediction. As another possible improvement, we can ignore terms only appearing infrequently in a class. One might also try to exploit PAN specific properties such as the requirement for equally distributed male/female problems and for the language variety 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.

References

1. Burrows, J.F. 2002. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *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., Potthast, M., & Stein, B. 2017. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: CLEF 2017 Labs and Workshops, Notebook Papers. *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.