An Analysis of Cross-Genre and In-Genre Performance for Author Profiling in Social Media

Maria Medvedeva(✉), Hessel Haagsma, and Malvina Nissim

University of Groningen, Groningen, The Netherlands
m.medvedeva@student.rug.nl, {hessel.haagsma,m.nissim}@rug.nl

Abstract. User profiling on social media data is normally done within a supervised setting. A typical feature of supervised models that are trained on data from a specific genre, is their limited portability to other genres. Cross-genre models were developed in the context of PAN 2016, where systems were trained on tweets, and tested on other non-tweet social media data. Did the model that achieved best results at this task got lucky or was it truly designed in a cross-genre manner, with features general enough to capture demographics beyond Twitter? We explore this question via a series of in-genre and cross-genre experiments on English and Spanish using the best performing system at PAN 2016, and discover that portability is successful to a certain extent, provided that the sub-genres involved are close enough. In such cases, it is also more beneficial to do cross-genre than in-genre modelling if the cross-genre setting can benefit from larger amounts of training data than those available in-genre.

Keywords: Author profiling · Cross-genre · Twitter · Blog · Social media

1 Introduction and Background

Promoting and evaluating research on user profiling, especially in social media, has been a key objective of the PAN evaluation Labs[1] in the past years [7–9]. The tasks have regularly attracted substantial numbers of participants, and by providing a variety of annotated datasets, have basically established the benchmark for author profiling in social media.

Social media, however, is a rather broad concept. It has been observed that at least in the context of user profiling, even within the realm of *social media* in general, not all data is the same. At the 2014 edition of the PAN Labs, the organisers provided four different types of widely defined social media data, namely Twitter, blogs, reviews, and an underspecified "social media" set to develop author profiling systems [8]. Training and testing was done *in-genre*, and average results in English are highest for Twitter, followed by blogs, reviews, and Social Media,

[1] pan.webis.de.

G.J.F. Jones et al. (Eds.): CLEF 2017, LNCS 10456, pp. 211–223, 2017.
DOI: 10.1007/978-3-319-65813-1_21

with performance differences up to 10% points. While the organisers suggested that the highest accuracies on Twitter might be due to the larger number of documents in that dataset [8] compared to blogs and reviews, the results might also be an indication that surface cues for predicting demographics are easier to grasp on Twitter than blogs, and that reviews are harder than either of them.

How much of such results depends on size and how much depends on genre is still an unanswered question, though it would be useful to have an answer. Assuming that we have very many tweets and not too many blogs, can we train our model on tweets and successfully test it on blogs, as both pertain to the *social media* wide genre? It is known that supervised models are pretty much bound to the training data they are exposed to. So, if social media data were more or less interchangeable in their sub-genres (e.g. Twitter, blogs), at least for profiling, one sub-genre dataset could serve for all and we could save a lot of annotation effort.

This kind of reflections, we believe, prompted the settings for the author profiling competition at PAN 2016, which was indeed organised in a *cross-genre* fashion [10]: while training data was provided in the form of user generated tweets, the test data was only known to be some kind of non-Twitter social media texts, with no other information available at system development time regarding its nature. After submission, the test set was eventually revealed to be blogs.

In building a system for this cross-genre task, we had aimed at exploiting features that might be typical for social media texts in general and focusing on language characteristics rather than metadata or genre-specifics (e.g. mentions, hashtags, etc.) [4]. This strategy proved successful, yielding indeed the best overall performance at the PAN 2016 cross-genre profiling task [10], training on tweets, and (unknowingly) testing on blogs. But beside the specifics of the PAN 2016 competition, have we really succeeded in developing a robust cross-genre model for social media? Operationally, in the context of this paper, we translate this question into the following two research questions:

Q1: is the Twitter-trained model that won the PAN 2016 cross-genre task on author profiling truly cross-genre so that good results can be observed in datasets other than the PAN 2016 test set?

Q2: if the features truly capture some general aspects of demographics, can the model be trained on datasets other than Twitter and still yield a good performance?

In order to provide an answer to Q1, we test our models on existing non-Twitter datasets from PAN 2014, and compare its performance to *in-genre* results in two ways, by cross-validation and by comparison to the official results of the PAN 2014 competition. To answer Q2, we train our model on existing non-Twitter datasets and test it on the same sets we test the Twitter-model on, still in a *cross-genre* setting. All these experiments are described in Sect. 3, results are provided in Sect. 4, and discussed in depth in Sect. 5. The GronUP system, which we use to run all experiments, is briefly summarised in Sect. 2.

2 GronUP

GronUP is a Support Vector Machine classifier that performs age and gender classification. It was developed in the context of the PAN 2016 author profiling competition, where it achieved best results. We offer here a short overview of its main features, while a full system description, together with the details of its performance at PAN 2016, can be found in [4].

2.1 Preprocessing

The first module in our system preprocesses the data. In the case of tweets, we use the tweets as is. For blogs, we split the document into sentences, in order to convert them into a similar structure as the tweets. That is, we treat tweets as being rough equivalents of sentences in other data types. The items (tweets/sentences) are then tokenised, using the Natural Language Toolkit (NLTK, [2]) TweetTokenizer.[2] Before tokenisation, some tokens are converted to placeholders: all URLs are replaced with the string 'URL', and all numbers are replaced by 'NUMBER'. Additionally, any HTML markup is deleted.

2.2 Classifier

For developing our system we used the Python Scikit-learn machine learning library [5]. Building on previous work and insights from participating systems in previous editions of this task, we opted to train a Support Vector Machine (SVM) with a linear kernel. We shortly evaluated the effect of different parameter values on performance using cross-validation on training data, but as a general approach we did not want to tune the system on tweets too closely. Indeed, while we observed that increasing the value of the cost parameter C led to an improvement in performance, we suspected that allowing for fewer incorrect classifications in the training data would lead to overfitting and worse performance on data from a different genre. We therefore focused on feature engineering rather than SVM settings, and used default parameters ($C = 1.0$) for all our models.

2.3 Features

In coming up with features to model user gender and age, we relied both on previous work and our own intuitions.[3] From a language perspective, we have aimed at creating general models, without language-specific features. In this section, we describe all of the features of GronUP.

N-grams. N-gram-based features have proven to be highly useful indicators of various linguistic differences between authors [1,8]. We use 1- to 3-grams for

[2] Contrary to what was stated in [4], TweetTokenizer is used for blogs, not word_tokenizer.

[3] In this paper we do not investigate the contribution of individual features, but insights on this can be found in the detailed description of GronUP [4].

tokens, and 2- to 5-grams for characters. We also add 1- to 3-grams for part-of-speech, since previous work indicates that female writers on social media tend to use more pronouns while male writers use more articles and prepositions, independently of their age [11]. We assign PoS-tags using the TnT tagger [3] trained on the Penn Treebank for English and on the CoNLL 2002 data for Spanish.

Capitalisation. We incorporate three capitalisation-related phenomena, which we believe to be indicative of user age: we expect younger users to be less keen on using sentence-initial capitals, we expect younger users to be more lenient when it comes to the capitalisation of specific tokens, such as proper nouns, and we expect them to be more likely to either use only lower-case or only upper-case letters. The sentence-initial capitalisation feature is implemented as the percentage of sentences (≈tweets) which start with a capital letter. For capitalisation of tokens, implementing a check for the correct application of language-specific capitalisation rules is too time-consuming and to language-specific for our purposes. Hence, we add this feature as the user-level average proportion of capitalised words in a sentence w.r.t. the total number of words in that sentence. The final capitalization feature is represented by the proportion of capital letters in a sentence as the total number of characters in that sentence. The resulting feature is the average of this proportion for a user, across documents.

Punctuation. The second group of orthographic features captures punctuation. We expect younger users to use 'proper' sentence-final punctuation less often, and to be more likely to use either very long sequences of punctuation (e.g. 'Nice!!!!!!!!!') or no punctuation at all. The first feature is represented as the percentage of items which have ".", "!" or "?" as their final token, the latter as the average proportion of punctuation characters w.r.t. all characters, across documents, for each user.

Average Word Length and Average Sentence Length. The use of longer words and sentences can be an indicator of a more advanced writing style, and it has been observed that older users tend to use longer words and sentences. We represent this characteristic as the average word length in characters and the average sentence/tweet length in tokens, per user.

Out of Dictionary Words. We hypothesise that the number of typos and slang or non-standard words can be a useful feature for distinguishing different age groups. For this feature we calculate the percentage of misspelled and out-of-dictionary words out of the total number of words per user. To detect out of dictionary words in English and Spanish we use the Python Enchant Library[4] [6] with `aspell` dictionaries for all available dialects.

Vocabulary Richness. The level of variety of a person's language use, represented here by the richness of their vocabulary, is a salient characteristic of their writing style. A way to measure the vocabulary richness of a person is to count how many words used by the user were only used once. The more unique words

[4] http://abisource.com/projects/enchant.

written by a user, the richer their vocabulary. We represent vocabulary richness as the percentage of words used only once w.r.t. to the total number of words used by an author.

High-Frequency Words. Highly frequent have been shown to be a key feature for determining the gender and age of an author [11]. Highly frequent, here, means those words which are used more frequently by users in one category than by users in other categories. Examples include sports-related words for male users versus female users, or popular textisms for the youngest users versus older users.

$$rf_{tc} = \frac{tf_{tc}}{tf_{t\neg c}} \tag{1}$$

The relative frequency (rf) of a term was calculated as in Eq. 1, where t denotes the term and c the class. We then used a ranked list of the top 2500 most frequent words for each category. The feature itself is a vector containing the occurrences of function words for each category in a document.

Emoticons. It has been found that, on American Twitter, younger people are more likely to use emoticons without noses (e.g. ':)') than older people, who prefer emoticons with noses (e.g. ':-)') [12]. Also, there is a correlation between the overall frequency of emoticon use, usage of non-nose emoticons, and age. As such, we implement features capturing the proportion of emoticons with noses out of all emoticons, and the proportion of emoticon tokens out of all tokens. In addition, based on our own intuitions about emoticon use among different user groups, we add features capturing the percentage of reverse emoticons (e.g. '(:') out of total emoticons, and the proportion of happy emoticons out of total emoticons.

Second-Order Attributes. Given that the goal of our model is to be able to generalise author profiling from Twitter to a different social media domain, we want to have feature representations that are not too closely tailored to the training data. For example, sentence length should ideally be represented as a relative value, indicating whether users from a certain class write longer sentences, on average, than users from another class. This is especially the case here, since tweets are not usually written as full sentences but other social media can be, and absolute values thus might not be have any meaning outside of the Twitter domain.

We implement second-order attribute representation of features to deal with this problem, similar to the approach in [1]. For the real-valued features (i.e. all except n-grams, function words), the mean is calculated for the training data, and the relative distance of the classes in each category is determined. Applying this to test data, the scores for a user are compared to the mean for the test data, and the difference is then compared to those found for the training data. If, in training, it is found that female users use five percent less capitals than the mean, and an unseen author gets the same relative score compared to the mean for the test data, the feature vector would indicate that the unseen author is more likely to be female than male.

3 Experiments

3.1 Datasets

We use several datasets to test the cross-genre robustness and portability of our system. Although GronUP at PAN 2016 was run on English, Spanish, and Dutch, the other datasets we use do not have Dutch data (and in one case only English data), so that we run experiments on English and Spanish only.[5] An overview of genres, source, and size in terms of users and total tokens is provided for each set in Table 1.

Table 1. Size of training sets (number of tokens and users) used in this paper. (*) We only use a portion of the original dataset of Social Media from PAN 2014 as it is far too large to use as such for training to carry out proper comparisons.

Genre	Source	ID	English tokens	Spanish tokens	English users	Spanish users
Twitter	PAN 16 training set	T2016	4306 K	1876 K	436	250
Blogs	PAN 14 training set	B2014	653 K	770 K	147	88
Social Media	PAN 14 training set(*)	S2014	23551 K	8494 K	1381	1271
Reviews	PAN 14 training set	R2014	1128 K	–	4160	–

All of the sets that we use for training and for testing in our experiments are *training sets* that come from the PAN competitions, since *test sets* are usually not available outside of the live shared task. Note that we want to profile authors based on their textual content, rather than using metadata, which is not included in the datasets we use.

Twitter. The portion of the training data that we use consists of tweets of users in English (436 users) and Spanish (250 users). We do not include Dutch. Also, we do not use the PAN 2014 tweets distribution because that dataset is almost entirely (96%) subsumed in the PAN 2016 Tweet training set.

Blogs. We use PAN 2014 blogs dataset, consisting of 147 manually annotated blogs in English and 88 in Spanish with up to 25 posts per blog. For the PAN 2016 evaluation the combination of training and test blog data of 2014 has been used (B2016-test).

Reviews. For the 'reviews' genre we have used hotel reviews dataset that have also been a part of PAN 2014, containing 4160 users and their reviews on TripAdvisor.[6]

[5] The choice for these languages is due to the availability of data, but the model could be trained on any language for which preprocessing tools (tokenizer, PoS-tagging, dictionary) and training data are available.

[6] http://www.tripadvisor.com.

Social Media. We have also used the 'social media' dataset distributed at PAN 2014 (non-Twitter, non-blogs, but not otherwise defined), which is a portion of the PAN 2013 corpus. In order to control for the size of the training data in the experiments, we took a smaller random sample of the Social Media dataset and reduced it to the size of Twitter PAN 2016, which still resulted in higher number of users and tokens (see. Table 2); the data in Spanish was of comparable size. The distribution of gender and age groups remained very similar to the original.

While gender is completely balanced in all datasets that we use for training with a 50/50 distribution of male and female users in all sets, the age groups are not so evenly distributed, with dominant classes being 35–49 and 25–34 in both tweets and blogs.

3.2 Experimental Settings

The experiments we run are aimed at targeting and answering the two research questions put forward in Sect. 1. We briefly report them again, and explain which experiments will answer which question, and how. For the sake of clarity, we refer to the specific lines in Tables 2 and 3 that provide experimental evidence towards each answer.

Q1: Is the Twitter-trained GronUP model truly cross-genre so that good results can be observed in datasets other than the PAN 2016 test set?

To answer this question, we need a measure of what "good performance" means on test sets other than the PAN 2016 blogs. We compare our cross-genre performance on other datasets (lines #6–#8) to new in-genre results, obtained in two different ways:

- Cross-validation on the same datasets the Twitter-trained model is tested on (lines #2–#4).
- The scores obtained by the systems officially submitted to the PAN 2014 competition, on the same datasets (lines #13–#15)

Two of the datasets that we consider (reviews and blogs, see Table 1) are smaller in size than the Twitter dataset, so cross-validation should yield, at least due to training data size, lower results than the cross-genre setting, unless the benefit of training and testing in-genre contributes highly. It should also be remembered that the Twitter model we submitted could not have been tuned to the test set, as nothing about test data was known to us, but the good performance could have still been chance.

Q2: If the features truly capture some general aspects of demographics, can the model be trained on datasets other than Twitter and still yield a good performance on different test sets?

Table 2. Cross-genre and within-genre results on profiling age and gender in Spanish (es) and English (en), for the various settings described in Sect. 3.2. For dataset abbreviations, please refer to Table 1.

#	Training set	Test set	Age (en)	Gender (en)	Age (es)	Gender (es)	Average
1	T2016	*(cross-val)*	0.4573	0.7067	0.4899	0.7085	0.5906
2	B2014	*(cross-val)*	0.3810	0.7143	0.4091	0.6590	0.5409
3	R2014	*(cross-val)*	0.3236	0.6526	–	–	*0.4881*
4	S2014	*(cross-val)*	0.3277	0.4946	0.3855	0.5951	0.4507
5	T2016	B2016-test	0.5897	0.6410	0.5179	0.7143	0.6157
6	T2016	B2014	0.5374	0.7347	0.4205	0.6818	0.5936
7	T2016	R2014	0.2377	0.5000	–	–	*0.3689*
8	T2016	S2014	0.3273	0.4960	0.3097	0.5628	0.4240
9	S2014	B2014	0.4354	0.5714	0.4318	0.5795	0.5045
10	S2014	R2014	0.2413	0.5115	–	–	*0.3764*
11	S2014	T2016	0.3601	0.5367	0.3985	0.5188	0.4535

Table 3. Best results from PAN 2014, per sub-task (i.e. not necessarily by the same participant). Scores are obtained from the official report [8].

#	Training set	Test set	Age (en)	Gender (en)	Age (es)	Gender (es)	Average
12	T2014	T2014-test	0.5065	0.7338	0.6111	0.6556	0.6268
13	B2014	B2014-test	0.4615	0.6795	0.4821	0.5893	0.5531
14	R2014	R2014-test	0.3502	0.7259	–	–	*0.5381*
15	S2014	S2014-test	0.3652	0.5421	0.4894	0.6837	0.5201

To answer Q2, we train our model on the Social Media dataset rather than Twitter and test it on the same sets we test the Twitter-model on, still in a *cross-genre* setting. As mentioned in Sect. 3.1, we subsampled the Social Media dataset which was originally much larger than the Twitter dataset to control for size of training data. We do not repeat this experiment on reviews/blogs as these datasets are too small in terms of tokens. Evidence towards answering this question will be provided by comparing lines #6–#8 with lines #9–#11 in Table 2.

Both age and gender predictions are structured as classification tasks, where gender is a two class problem, and age values are binned into five separate classes. In all experiments we use GronUP for classification.

4 Results

The results of the experiments described in Sect. 3 are shown in Table 2. In what follows, we refer to each setting by means of its line number in the table.

We distinguish three types of experiments: in-genre using cross-validation (#1–#4), cross-genre, training on tweets (#6–#8), and cross-genre, training on Social Media (#9–#11). Each result is represented by the accuracy score on each subtask (English/Spanish & age/gender), and the average over subtasks. The exception here are the results on review data, which do not include Spanish (#3, # 7, and #10). Generally, there is little difference between the performance on each subtask and in the overall average. That is, when the system does better in one setting than another, it also tends to do better on each of the subtasks in the one setting. This makes the results easier to interpret, since we can focus mainly on the average accuracy score, when comparing performances between settings.

Note that we also report GronUP's results for the official setting of PAN 2016 (#5, [4,10]). This data is not available to use in any of the other experiments, so we cannot compare these results directly to the new results reported in this work. However, the test set of PAN 2016 is a superset of the PAN 2014 blogs training data, and we can use the latter as a proxy for the PAN 2016 test set. This is corroborated by the fact that the performance on the PAN 2016 test set (#5, 0.6157) is very similar to the performance on the PAN 2014 blogs training set (#6, 0.5936), using the same training data.

By comparing the system's performance in different settings, we hope to answer the research questions posited in Sect. 3.2. First, we compare the results in cross-validated in-genre settings to the cross-genre settings where we train on Twitter data. We see that results are mixed: on blog data, cross-genre performance (#6, 0.5936) is surprisingly higher than in-genre performance (#2, 0.5409). On review data, on the other hand, the system performs clearly worse in the cross-genre setting (#7, 0.3689) than in the cross-validation setting (#3, 0.4881). On Social Media, cross-genre performance (#8, 0.4240) is also lower than in-genre (#4, 0.4507), but the difference is much smaller, with a drop of only 2.5% points, compared to 12% points for reviews.

As a second point of comparison, we use the best performances in the PAN 2014 shared tasks. The highest accuracy score on each subtask for PAN 2014 is displayed in Table 3. Note that these results are on the test sets of PAN 2014, whereas the results in Table 2 are on the cross-validated training sets of PAN 2014, since we do not have access to the test data. Still, we can assume that these sets are drawn from the same underlying pool of data, and thus can make valid comparisons. We can compare these state-of-the-art results to both the cross-validated in-genre performance and the cross-genre Twitter-trained performance of GronUP. For blog data, we see that our system performs very similarly to the best systems of 2014 when cross-validating (#2, 0.5409 vs. #13, 0.5531) and does even better in the cross-genre setting (#6, 0.5936). On review data, the picture is the complete opposite. In-genre, we see lower results (#3, 0.4881 vs. #14, 0.5381), and cross-genre, they are a lot worse (#7, 0.3689). Social Media shows a similar effect, with an in-genre difference of 7% points (#4, 0.4507 vs. #15, 0.5201), and a cross-genre difference of 10% poings (#8, 0.4240 vs. #14, 0.5201).

Lastly, we assess whether GronUP is dependent on having tweets as training data, or whether it can generalise to having different social media text as training data. For review data, this seems to be the case, as performance is low, but similar, in both the Twitter-trained setting (#7, 0.3689) and the social-media-trained setting (#10, 0.3764). On blog data, there is a clear drop, from 0.5936 (#6) when trained on Twitter, to 0.5045 (#9) when trained on Social Media. However, given that the within-genre state-of-the-art on blogs is 0.5531 (#13), results in both settings are far from poor.

Additional insight can be gained by looking at the same cross-genre setting, with different directionality, as in #8 and #11. Here, accuracy is somewhat lower when going from Twitter to Social Media (#8, 0.4240) than when going from Social Media to Twitter (#11, 0.4535), but the difference between cross-validation on tweets (#1, 0.5906) and cross-validation on Social Media (#4, 0.4507) is much larger.

5 Analysis

Results appear to provide a rather mixed picture in terms of performance, but we believe they can be explained according to three aspects, namely *size of training data*, *gap in genre*, and *quality of data*.

The fact that a Twitter-trained model performs better on blogs than cross-validating on blogs (#6 vs. #2) can be explained with a substantial difference in size in training sets, and the cross-genre potential loss is limited because of a narrow genre-gap. That is, we hypothesise that tweets and blogs are relatively similar, especially topic-wise. Conversely, we expect the difference between tweets and reviews to be much larger; the main source of difference, in addition to writing style, being that the content of tweets (and blogs) mainly concerns the author themselves, whereas reviews are mostly about the thing that is being reviewed.

This, indeed, explains why training on tweets and testing on reviews (#7) yields a significant drop in performance, when compared to cross-validated results on reviews (#3), official results from PAN 2014 (#14), and also with respect to testing on blogs (#6), which are a domain closer to training data than reviews. Therefore, we can speculate that, in a cross-genre setting, the main influencing factor is the difference in genre, and that, if genres are sufficiently similar, an increase in the amount of training data can further boost performance. This is corroborated by the performance of the social media-trained model (#9–#10), which is also clearly better on blogs than on reviews (assuming the genre-gaps are very similar to those for Twitter).

Such a hypothesis makes sense theoretically and is supported by the data, but leaves us with one additional question: why does GronUP perform about 10% points worse when going from Social Media to blogs (#9) than when going from Twitter to the same blogs (#6)? Similarly, cross-validation on Social Media (#4) is a lot less accurate than on Twitter (#1). The genre-gap we discussed above cannot explain this, since we assess tweets, blogs, and social media to be

all quite similar to each other, and distinct from reviews. One possibility is that GronUP has limited portability: it manages to create a model of age and gender from Twitter data, for which it was developed, but this does not extend to other types of training data, even similar ones like social media.

This is possible, but another explanation seems more likely: based on manual inspection of the social media data, we believe that the performance difference is due to the relatively lower quality of the Social Media data. This hypothesis is supported by the fact that Social Media results are the lowest in the PAN 2014 official results (#15) in spite of this dataset being the largest of all in that competition. At PAN 2014, moreover, models were developed in-genre, and therefore should have been more capable of utilising the social media data for training.

We would like to conclude the analysis of the results with an observation regarding GronUP's in-genre performance overall. In [4], it was suggested that "*We believe that our lower scores [w.r.t. the in-genre state-of-the-art] are likely due to the conscious choice to avoid the use of potentially strong age/gender indicators which would work on Twitter only.*" The additional results seem to confirm this idea, although the difference between GronUP's in-genre performance (#1–#4) and the state-of-the-art (#12–#15) is not excessively big. On tweets, the difference is 4% points, on blogs 1, on reviews 5, and 7 on Social Media. In making such comparison, it should also be noted that our scores are on the cross-validated training sets of PAN 2014, while the official results reported in Table 3 are based on training on the whole set (thus slightly larger than what we used) and testing on separate data.

6 Conclusions

We set out to investigate the robustness of our cross-genre profiling system, beyond the good performance obtained at the PAN 2016 author profiling task. We did so via a series of experiments that revolved around two main research questions, and involved training and testing models on datasets pertaining to different sub-genres under a more general *social media* label. We ran in- and cross-genre evaluations, compared results across the various settings, and also against official results that had been obtained on the same datasets at the PAN 2014 evaluation campaign. How then do the obtained results and our analysis answer our research questions?

In Q1, we were asking if the GronUP Twitter-model would perform well beyond the PAN 2016 dataset. The answer is yes, it does, as long as the genre-gap is not too broad. Good performance on blogs, and poorer performance on reviews, provide support for this claim. Additionally, a general advice that we glean from the experiments from a practical perspective, is that, provided the genres are similar enough, size matters, and a large cross-genre training set can be more useful than a small in-genre one.

In Q2, we were asking whether GronUP's features do indeed capture demographics beyond Twitter, and could therefore be used to build satisfactory models on training sets of different genres. We trained a profiling models on the

PAN 2014 Social Media data, and observed that results are completely comparable to those obtained when training on Twitter when testing on reviews, but are quite almost 10 point lower when testing on blogs. The answer is not definite, but the picture is particularly blurred by the quality of the Social Media data, which makes this set rather unreliable to draw any safe conclusions.

As a byproduct of our investigation, we believe we can also provide a preliminary answer to the question the organisers of PAN 2014 put forward when observing the discrepancy in in-genre performance between tweets (higher), blogs (lower), and reviews (even lower) [8]. As mentioned in Sect. 1 of this paper, they had wondered whether the gap was due to training size (larger for tweets and smaller for the other datasets) or some intrinsic complexity of the data itself, leaving this issue to further investigation. Our analysis seems to point towards the size explanation, as we have seen that cross-validating on blogs yields a worse performance than training on tweets and testing on blogs. Some more specific data characteristic appears to play a role in cross-genre settings.

This last point is important with respect to the direction the development of manually annotated training data should take (if the genres are similar enough, can we just concentrate on one and build a single very large training set?), and definitely deserves further investigation. Indeed, in future work, we would very much like to better understand and model the trade-off between these two crucial aspects: size and genre-gap. While the first one is straightforward to quantify, the second one will require the development of some more complex, dedicated measure. Moreover, we plan to explore the contribution of additional datasets, especially to provide better evidence in answering Q2, potentially also in different experimental settings, so as to test more robustly the benefits and drawbacks on cross-genre profiling.

References

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Jair-Escalante, H.: INAOE's participation at PAN'15: author profiling task. In: Proceedings of CLEF (2015)
2. Bird, S., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol (2009)
3. Brants, T.: TnT: a statistical part-of-speech tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing, pp. 224–231. Association for Computational Linguistics (2000)
4. Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M.: GronUP: Groningen user profiling. In: Working Notes of CLEF, pp. 846–857. CEUR Workshop Proceedings. CEUR-WS.org (2016)
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
6. Perkins, J.: Python 3 Text Processing with NLTK 3 Cookbook. Packt Publishing Ltd. (2014)

7. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF (2015)
8. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the author profiling task at PAN 2014. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) Working Notes for CLEF 2014 Conference, Sheffield, 15–18 September 2014. CEUR Workshop Proceedings, vol. 1180, pp. 898–927. CEUR-WS.org (2014)
9. Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation, pp. 352–365. CELCT (2013)
10. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2016
11. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging, vol. SS-06-03, pp. 191–197. AAAI Press (2006)
12. Schnoebelen, T.: Do you smile with your nose? Stylistic variation in Twitter emoticons. Univ. Pa. Work. Pap. Linguist. **18**(2), 117–125 (2012)