Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016

Notebook for PAN at CLEF 2016

Pashutan Modaresi, Matthias Liebeck, and Stefan Conrad

Institute of Computer Science
Heinrich Heine University Düsseldorf
D-40225 Düsseldorf, Germany
{modaresi,liebeck,conrad}@cs.uni-duesseldorf.de

Abstract Author profiling deals with the study of various profile dimensions of an author such as age and gender. This work describes our methodology proposed for the task of cross-genre author profiling at PAN 2016. We address gender and age prediction as a classification task and approach this problem by extracting stylistic and lexical features for training a logistic regression model. Furthermore, we report the effects of our cross-genre machine learning approach for the author profiling task. With our approach, we achieved the first place for gender detection in English and tied for second place in terms of joint accuracy. For Spanish, we tied for first place.

1 Introduction

Author profiling deals with the study of various profile dimensions of an author [2]. The focus of this study is to gain an understanding of how authors of different classes (e.g., old men and young women) use different characteristics while writing text and which textual features might be characteristic for all people in the same class. For instance, younger people might make more spelling mistakes than older people.

Due to its applications in fields such as security, forensics and marketing, the study of various profile aspects of an author has gained more importance in recent years. This in turn has attracted the attention of the scientific community [1]. More specifically, the PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse) competition has been focusing on the task of author profiling as a part of the CLEF conference since 2013.

Author profiling is useful in a context where missing information about authors is relevant for an organization. For instance, a company might want to know how old their target group in social media is in order to customize advertising campaigns. In other contexts, such as *political online participation* [8], where cities allow their citizens to participate in politics via internet, it is interesting to automatically estimate demographic distributions of the users without the need to directly ask them for personalized data. Even in fields such as abstractive text summarization, author profiling techniques can be used to differentiate between human-written and machine-generated summaries [9].

Two profile aspects, namely *age* and *gender*, have been the focus of the PAN author profiling competitions. The focus of the 2016 shared task [15] is on cross-genre age and gender identification. That means that the training documents are on one genre (Twitter) and the evaluation is on another (unknown to the participant at the time of the software submission) genre, such as blogs or social media. English, Spanish and Dutch are the languages that were addressed in this year's challenge.

2 Related Work

Author profiling has been a recurrent PAN task since 2013. Until today, the age and gender classification tasks have always been part of the author profiling challenge. The first challenge in 2013 [14] was on English and Spanish blog posts. The focus in 2014 [13] was on four domains (*blogs*, *Twitter*, *social media*, and *hotel reviews*), each of which was provided with an individual training and test set. The 2015 challenge [12] was on English, Spanish, Italian, and Dutch tweets and provided an additional classification task of identifying personality traits (extroversion, emotional stability, agreeableness, conscientiousness, and openness to experience). The challenge in 2016 also consisted of English, Spanish, and Dutch tweets as training data but had the additional difficulty of a cross-genre evaluation dataset.

Since there have been 53 participating teams in the last 3 years, various approaches to author profiling have been tested. The teams have used different preprocessing steps, features and classifiers. [12] provides an overview of the approaches of the participating teams in the 2015 challenge: For preprocessing, steps such as removing HTML code, removing hashtags and URLs, lowercasing text, and stop word filtering have been used. Character n-grams, word n-grams, POS n-grams, punctuation signs, topic modeling with Latent Semantic Analysis (LDA), and Twitter-specific features, such as links, hashtags, and mentions, have frequently been used as features. The most frequently trained classifier is the support vector machine.

In our approach, we need to keep in mind that the evaluation is cross-genre. This means that we cannot use features that are specific for Twitter, such as hashtags. Furthermore, we have to take into account that tweets are limited to 140 characters in length and our evaluation genre may be comprised of longer text. Features based on absolute length, like word counts, were, therefore, not relevant in this year's challenge. As a result, all of our features are normalized to account for the domain change.

Furthermore, there has also been research into author profiling outside of PAN, for instance, to predict demographic information [16], such as annual income, having children, religious beliefs, and education levels from Twitter users.

3 Methodology

This section describes our approach to this year's PAN Author Profiling challenge. First, we outline preprocessing steps that we used to clean the data. Then, we describe the features that we used in our machine learning approach. Afterwards, we briefly explain why we chose logistic regression as our classifier. The dockerized source code of our profiler is available on GitHub[1].

3.1 Preprocessing

As the genre of the training and test sets are not the same, we processed the documents in the training set in such a way that most of the genre-specific information was eliminated. In this way, the risk of overfitting on genres other than *Twitter* was reduced. This was accomplished by a composition of multiple preprocessors that each map an input document d (string of characters) to a modified document d'. The individual preprocessors are defined as follows:

- $p_1(d)$: Returns a string in which all case-based characters have been lowercased.
- $p_2(d)$: Filters all occurrences of URLs in the string. This is an important step toward creating genre-neutral documents.
- $p_3(d)$: A mention is a tweet that contains another user's @username anywhere in the body of the tweet and does not occur in other genres. This function eliminates all mentions in the document.
- $p_4(d)$: Hashtags are used to categorize tweets. Although hashtags may contain important information about the profile of an author, obtaining a meaningful representation of it is not always trivial (e.g., #timetoact). This function eliminates all hashtags from the document.
- $p_5(d)$: A retweet is a re-posting of someone else's tweet. As this feature is also tweet-specific and may not generalize to other domains, we eliminate all retweets.
- $p_6(d)$: Although the training set is claimed to be only consisting of English, Spanish, and Dutch tweets, it also contains tweets in other languages such as Arabic and Persian. We consider these tweets as noise by eliminating all non-latin characters from the input document.
- $p_7(d)$: Specific lexical features such as unigrams and bigrams result in better accuracies when accents are removed from the input document. This is accomplished by means of this function.
- $p_8(d)$: Eliminates all non-alphabetic characters from an input document. This function is applied when dealing with token-based features.
- $p_9(d)$: Eliminates all stop-words from the document using a language-specific predefined list.

The composition of preprocessors is feature-specific, meaning that for each feature a distinct set of functions is composed in order to preprocess the document (A detailed description of features is provided in Section 3.2). Table 1 lists the preprocessors per feature (category).

3.2 Features

After preprocessing the tweets, we need to extract features for a vector representation. The challenge in our particular task is the need for features that are genre-independent.

[1] https://github.com/pan-webis-de

Table 1. Features and their corresponding preprocessors

Feature	Preprocessors
Unigrams	$(p_9 \circ p_8 \circ p_7 \circ p_6 \circ p_5 \circ p_4 \circ p_3 \circ p_2 \circ p_1)(d)$
Bigrams	$(p_8 \circ p_7 \circ p_6 \circ p_5 \circ p_4 \circ p_3 \circ p_2 \circ p_1)(d)$
Average Spelling Error	—-
Character N-Grams	$(p_2 \circ p_3 \circ p_4 \circ p_1)(d)$
Punctuation Features	—-

Even though we use tweets from Twitter for training our classifier, we cannot use features that are specific for Twitter because the evaluation dataset is from another domain. Features that depend on absolute text length are not good candidates because tweets are limited to 140 characters, whereas the text length in the evaluation domain is probably unrestricted.

Since the tweets are in three different languages, we can either find language-specific features or language-independent features. Given that we are not familiar with all languages in this task, we decided to find language-independent features.

We experimented with multiple features in the course of our experiments. In the end, we decided to use the combination of features that worked best on the *Blog* dataset from 2014 as test set (with the dataset from 2016 as training set):

- **Word unigrams** that occur at least two times
- **Word bigrams**
- **Character 4-grams** within word boundaries
- We utilize Hunspell[2] with LibreOffice dictionaries for all three languages to measure an **average spelling error** by determining a relative value for correctly spelled words.
- In addition, we make use of four token-based **punctuation features**: average comma count, average dot count, average exclamation count, and average question mark count.

All these features were used for the *age* subtask. We omitted the punctuation features for the *gender* subtask.

We also experimented with punctuation n-grams and used *polyglot*[3] to retrieve L_2-normalized POS-Tag distributions of UTS tags [11]. Unfortunately, both attempts only worsened our results and we did not pursue them further.

3.3 Classification

We used logistic regression [10] to train our final models. Logistic regression belongs to a family of classifiers that have high bias and low variance. Although this classifier has a low variance (which could lead to underfitting), due to the cross-genre nature of the problem, this will not have negative implications as the test dataset does not consist of tweets. On the other hand, logistic regression has a high bias which can lead to overfitting. This could also be handled using regularization techniques. These

[2] http://hunspell.github.io/

[3] http://polyglot-nlp.com/

properties make the logistic regression classifier a suitable choice for our cross-genre classification problem.

As the classification task is a multiclass problem, we use the one-vs-rest scheme for logistic regression. Moreover, we set $C = 10^{-3}$ as the regularization strength.

Additionally, we also experimented with random forest [3] and gradient boosting [5]. Both of these techniques use randomization to build decision trees (or regression trees) to combat overfitting. The results obtained using logistic regression were superior to both above-mentioned classifiers.

4 Evaluation

As the focus of this year's competition is cross-genre author profiling, we only used the tweets dataset provided by the organizers to train our models. For *age*, the following classes are provided: 18-24, 25-34, 35-49, 50-64 and 65-xx. Moreover, *gender* consists of the two classes: *male* and *female*.

The provided dataset is in the three languages English, Spanish and Dutch. The corpus was annotated with the age and gender information of the authors, except for Dutch, which was only annotated with gender information. For each individual author, there exists an XML document consisting of several tweets. For English there are 436 documents, for Spanish 250 documents, and for Dutch 384 documents. In our approach, we concatenated all tweets of an author into a single document.

In order to evaluate our models, we used stratified k-fold cross validation $(k = 10)$ on the tweets dataset. For this we used the implementation provided by *scikit* [4]. Furthermore, we used the available training datasets from PAN2014 to evaluate our models on genres other than Twitter (blogs, social media, reviews). This was accomplished by using the TIRA experimentation platform, which provides a service to handle software submissions [6, 7].

4.1 Official PAN 2016 Benchmark

For each language (except Dutch, which contained only the age annotation), two distinct models were trained: one for gender and one for age. For both labels we used the same set of features, except for punctuation features that were only used for age.

For the final evaluation, two test datasets were provided by the task organizers, where the first dataset is a subset of the second one. The official results for the first and second test datasets are reported in Table 2 and Table 3 accordingly.

Table 2. Evaluation results in terms of accuracy for the first test dataset

Language	Joint	Gender	Age
English	0.1552	0.5029	0.3017
Spanish	0.2031	0.6406	0.2813
Dutch	0.5	0.5	—-

Table 3. Evaluation results in terms of accuracy for the second test dataset

Language	Joint	Gender	Age
English	0.3846	0.7564	0.5128
Spanish	0.4286	0.6964	0.5179
Dutch	0.5040	0.5040	—-

In general, higher accuracies are achieved on the second dataset. On the second test dataset, for both English and Spanish, the accuracies for age prediction are slightly above 0.5. The highest accuracy (0.7564) is achieved by gender classification for the English language. Outstanding is the joint accuracy of 0.4286 for the Spanish language. Also, the lowest accuracies are reported for Dutch. Unfortunately, at the time of authoring this work, no access to the test datasets was granted to explain this behavior. As all features used for gender classification are token centric, we assume that the out-of-vocabulary rate is too high in the prediction phase and this leads to the unsatisfiable results for Dutch.

4.2 Cross-Genre Effects

One of the main intentions behind using simple features in our approach is to avoid overfitting on genres other than Twitter. We also performed tests on the training datasets from PAN 2014 which are publicly available. The training dataset of PAN 2014 is also annotated with age and gender information and both labels have exactly the same categories as in PAN 2016. In comparison with PAN 2016, the PAN 2014 corpus only contains English and Spanish documents belonging to four different genres, namely blogs, Twitter, social media and hotel reviews. The accuracies of our model on blogs and Twitter are reported in Table 4 and Table 5 respectively.

Table 4. Evaluation results in terms of accuracy for PAN2014 training dataset (Blogs)

Language	Joint	Gender	Age
English	0.3878	0.8435	0.4830
Spanish	0.4091	0.7727	0.4773

Table 5. Evaluation results in terms of accuracy for PAN2014 training dataset (Twitter)

Language	Joint	Gender	Age
English	0.9510	0.9804	0.9542
Spanish	0.4270	0.7640	0.5281

Among all genres, the highest joint accuracies are achieved for blogs with 0.3878 for English and 0.4091 for Spanish. These values are even higher than the ones obtained during k-fold cross validation on PAN 2016 tweet dataset and signal that no overfitting occurred in case of blogs. It can also be observed that the accuracies for English tweets are extremely high with a score above 0.9. This is most probably due to the high overlap between the datasets from 2014 and 2016.

Table 6. Evaluation results in terms of accuracy for PAN2014 training dataset (Socialmedia)

Language	Joint	Gender	Age
English	0.2000	0.6000	0.2000
Spanish	0.1627	0.5951	0.2563

Table 7. Evaluation results in terms of accuracy for PAN2014 training dataset (Reviews)

Language	Joint	Gender	Age
English	0.1524	0.6067	0.2572
Spanish	—-	—-	—-

Unlike in the case of blogs, the accuracies for the genres social media and reviews are not satisfactory (see Table 6 and Table 7). The lengths of the documents in social media and reviews are much greater than the length of the documents in Twitter. This leads to a high out-of-vocabulary rate and consequently to unsatisfactory results.

Table 8. Confusion matrix for PAN2014 (English/Blogs/Age)

		Predicted					
		18-24	25-34	35-49	50-64	65-xx	∑
Actual	18-24	0	3	3	0	0	6
	25-34	0	28	32	0	0	60
	35-49	0	9	45	0	0	54
	50-64	0	1	20	2	0	23
	65-xx	0	0	3	1	0	4
	∑	0	41	103	3	0	147

As the measure of accuracy is not suitable to study the performance of our approach for each individual category, we also exemplarily provide the confusion matrix for a model trained on PAN 2016 tweets and tested on PAN 2014 English blogs on the category age (see Table 8). The confusion matrix shows that no instance from the categories 18-24 and 65-xx is correctly classified. The reason for this is the low support of these categories, which implies that the classifier has not enough data to learn from. Another interesting point is the high similarity between the categories 25-34 and 35-49. From 60 instances in the category 25-34, 32 instances are incorrectly classified to the class 35-49. This indicates that the features defined in our approach are not capable of discriminating between the aforementioned categories.

5 Conclusion and Future Work

We have presented our approach for the cross-genre PAN 2016 author profiling task. Our best results for the gender and age classification tasks in terms of accuracy are 0.7564 for English and 0.5179 for Spanish, respectively. Furthermore, we evaluated our approach on multiple genres to explore the effects of cross-genre machine learning.

The training set for age was imbalanced (see the confusion matrix in Table 8), which resulted in poor performance. We could use techniques such as sampling or SMOTE to tackle this problem.

In our experiments, we tested different feature combinations. It turned out to be difficult to find good genre- and language-independent features. For instance, the POS distribution turned out not to be a good genre-independent feature. In our future work, we will include more language-dependent features to better capture the characteristics of each language. Additionally, we will include lists of sentiment-bearing words in our features.

Acknowledgments

This work was partially funded by the PhD program *Online Participation*, supported by the North Rhine-Westphalian funding scheme *Fortschrittskollegs* and by the German

Federal Ministry of Economics and Technology under the ZIM program (Grant No. KF2846504).

References

1. Álvarez-Carmona, M.Á., López-Monroy, A.P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Escalante, H.J.: INAOE's Participation at PAN'15: Author Profiling task. In: CLEF (Working Notes). CEUR Workshop Proceedings, vol. 1391. CEUR-WS.org (2015)
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically Profiling the Author of an Anonymous Text. Commun. ACM 52(2), 119–123 (2009)
3. Breiman, L.: Random Forests. Mach. Learn. 45(1), 5–32 (2001)
4. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122 (2013)
5. Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29, 1189–1232 (2000)
6. Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). pp. 1125–1126. ACM (2012)
7. Gollub, T., Stein, B., Burrows, S., Hoppe, D.: TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In: 9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA. pp. 151–155. IEEE (2012)
8. Liebeck, M., Esau, K., Conrad, S.: What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld. In: Proceedings of the 3rd Workshop on Argument Mining. p. (in press). Association for Computational Linguistics (2016)
9. Modaresi, P., Conrad, S.: On definition of automatic text summarization. In: Proceedings of Second International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC2015). pp. 33–40 (2015)
10. Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. Journal of the Royal Statistical Society, Series A, General 135, 370–384 (1972)
11. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA) (2012)
12. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CLEF (2015)
13. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd Author Profiling Task at PAN 2014. In: Working Notes for CLEF 2014 Conference. pp. 898–927. CLEF (2014)
14. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Working Notes for CLEF 2013 Conference. CLEF (2013)
15. 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 (Sep 2016)
16. Volkova, S., Bachrach, Y.: On Predicting Sociodemographic Traits and Emotions from Communications in Social Networks and Their Implications to Online Self-Disclosure. Cyberpsychology, Behavior, and Social Networking 18(12), 726–736 (2015)