

# Cross-genre Age and Gender Identification in Social Media

Anam Zahid<sup>1</sup>, Aadarsh Sampath<sup>1</sup>, Anindya Dey<sup>1</sup>, Golnoosh Farnadi<sup>2,3</sup>

<sup>1</sup>Center for Data Science, University of Washington Tacoma, WA, USA

<sup>2</sup>Dept. of Appl. Math., Comp. Science and Statistics, Ghent University, Belgium

<sup>3</sup>Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium

**Abstract.** This paper<sup>1</sup> gives a brief description on the methods adopted for the task of author-profiling as part of the competition PAN 2016 [1]. Author profiling is the task of predicting the author’s age and gender from his/her writing. In this paper, we follow a two-level ensemble approach to tackle the cross-genre author profiling task where training documents and testing documents are from different genres. We use the soft-voting approach to build the classification ensemble. To include various feature sets, we first train logistic regression models using the extracted word n-gram, character n-gram, and part-of-speech n-gram features for each genre. We then ensemble single-genre predictive models trained on the blog, social media and Twitter data sources, to build our multi-genre ensemble approach. The experimental results indicate that our approach performs well in both single-genre and cross-genre author profiling tasks.

**Keywords:** Gender identification, Age prediction, Ensemble technique, Text mining, Cross-genre classification, Author profiling

## 1 Introduction

The rapid development of social media platforms has led to a massive volume of user-generated text in the form of blog posts, status updates, and tweets. This has generated great research interest in identifying authors’ profile [2]. Author profiling is the task of predicting the authors age and gender information with his/her writing. Most of the recent works in author profiling address the problem as a single-genre task where the instances of the training set and the test set are coming from a single platform. Due to the difficulties of gathering ground truth data for every platform, cross-genre author profiling task has been proposed. Cross-genre profiling has been done for personality prediction in [3], however little work has been done for identifying the age and gender of users in a cross-genre setting. Such models could be applied to environments where training data representative for the deployment domain is not available. Effective features from the recent works in age and gender classification were both content features such as unigrams, bigrams and word classes as well as stylistic features, such as part-of-speech (POS), slang words and average sentence length. For instance, in case of the gender identification, Villena Román et al. [4] extracted n-grams or bag-of-words as content features. In [5], Argamon et al. approached the task of gender identification by combining function words with POS tags. Given the related works in this domain, we include various feature sets in our model by training logistic regression models using the extracted word n-gram, character n-gram, and POS n-gram features from the documents. We propose a two-level ensemble approach which is a multi-genre predictive model

---

<sup>1</sup> This paper is an extended abstract

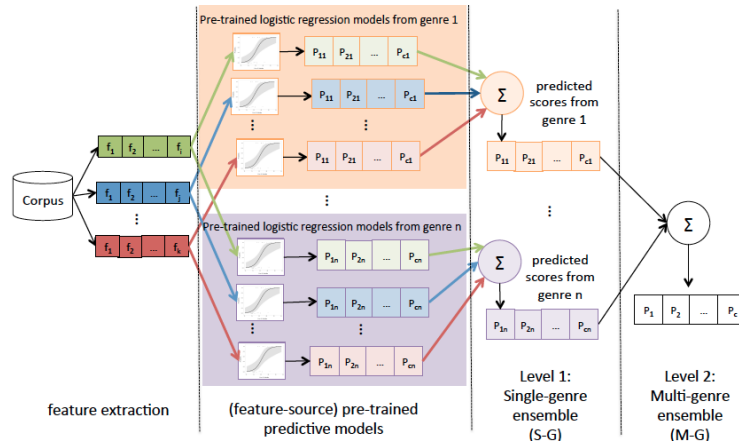


Fig. 1: The architecture of the multi-genre ensemble model .

that ensembles single-genre predictive models from the available ground-truth datasets of various genres, i.e., the blog, social media and Twitter datasets. Our multi-genre ensemble approach leverages various types of documents as training examples which makes it suitable for the cross-genre author profiling of the PAN2016 competition where the testing documents are from a hidden genre. The experimental results indicate that our ensemble approach can be used for both single-genre and cross-genre author profiling tasks. The rest of this paper describes the details of our submission to the PAN 2016 cross-genre author profiling task.

## 2 Methodology

Let us assume  $U$  is a set of all authors, where  $U = U_{train} \cup U_{test}$ . For all users in  $U_{train}$ , we know their age and gender, and our aim is to predict the age and gender of all users in  $U_{test}$  based on their written text. If  $U_{train}$  and  $U_{test}$  are coming from one platform (aka genre), we call the task a *single-genre author profiling* task, and if  $U_{train}$  and  $U_{test}$  are from different social media platforms, we call the task a *cross-genre author profiling* task. The overall architecture of our proposed ensemble approach for a single-genre (S-G) and multi-genre author profiling (M-G) is shown in Figure 1. Using the S-G ensemble approach, we incorporate various features extracted from the documents and by using the M-G ensemble approach, not only do we use different features, but also leverage predictive models of different genres which makes the framework suitable for cross-genre author profiling task.

**2.1 Pre-processing and data description:** The data provided by the PAN organizers, was in the form of XML documents from which user contents were extracted and cleaned by removing HTML tags and stop words. To tackle the cross-genre author profiling task, we collected data from 2014 and 2015 PAN author profiling contests and added them to our training dataset. For English and Spanish, we made three datasets from different genres: (1) *social media* with 7,746 documents for English and 1,272 documents for Spanish, (2) *blog* with 147 documents for English and 88 documents for Spanish and (3) *Twitter* with 576 documents for English and 340 documents for Spanish. For the Dutch dataset, we gathered data from *Twitter* with 418 documents. In all the datasets the

Table 1: Statistics of the combined datasets w.r.t. the users’ age.

Genre	Language	[18, 24]	[25, 34]	[35, 49]	[50, 64]	[65, xx]
blog	English	6	60	54	23	4
	Spanish	4	26	42	12	4
social Media	English	1,550	2,098	2,246	1,838	14
	Spanish	330	426	324	160	32
Twitter	English	86	200	204	80	6
	Spanish	38	110	148	38	6

gender distributions are uniform. The statistics of the combined datasets w.r.t. the frequencies of the five age groups (i.e., [18, 24], [25, 34], [35, 49], [50, 64], and [65, xx]) are shown in Table 1. Note that for the Dutch dataset we do not have the age of the authors.

**2.2 Feature extraction:** To create our feature space, we extract three different categories of features, drawing inspiration from related works. All the implementations are based on the machine learning package in Python called scikit-learn<sup>2</sup>. The extracted features are (1) *word n-gram* where  $n = \{1, 2, 3\}$  (aka uni, bi and tri-grams) using TF-IDF as a weighting mechanism, (2) *character n-gram* where  $n = \{3, 4, 5, 6, 7\}$  using TF-IDF as a weighting mechanism. To reduce the size of the feature space, we select  $k$  top features using Chi-square hypothesis testing where  $k = 5000$ , and (3) *POS n-gram*: in which we extract part-of-speech (POS) tags from each document using nltk package in Python<sup>3</sup>. Then each word in text is mapped to its corresponding POS tag and the text comprising of those POS tags is used to extract n-gram features with the same configuration of word n-gram with  $n = \{1, 2, 3\}$  and TF-IDF weighting.

Table 2: Accuracy of the age and gender prediction using single-genre ensemble (S-G) approach. Values in bold are higher than the majority baseline (base). All results are averaged over a 5-fold cross-validation

Genre	English		Spanish		Dutch					
	base	S-G	base	S-G	base	S-G				
blog	0.50	<b>0.65</b>	0.41	0.32	0.50	<b>0.66</b>	0.48	0.48	-	-
social media	0.50	<b>0.54</b>	0.29	<b>0.34</b>	0.50	<b>0.60</b>	0.33	<b>0.34</b>	-	-
Twitter	0.50	<b>0.60</b>	0.35	<b>0.46</b>	0.50	<b>0.57</b>	0.43	<b>0.48</b>	0.50	<b>0.53</b>

**2.3 Predictive model:** We train binary classifiers for predicting the gender of users and multi-class classifiers for predicting their age. For age and gender prediction tasks, we train three predictive models using the three feature sets that we explained above with logistic regression as a classifier for each genre-label-language. We then apply an *ensemble soft-voting* approach using the prediction scores of the models. The results of applying our S-G ensemble approach on the Twitter, social media and blog datasets are presented in Table 2. Our S-G ensemble approach outperforms the majority baseline in predicting the gender of users for all the three datasets for all three languages, however for the task of age prediction, our approach outperforms the baseline for the social media and Twitter datasets for English and Spanish. To tackle the cross-genre author profiling task, we first made S-G ensemble models for each genre, e.g., regarding the English dataset, we made three S-G ensemble models for the social media, blog and Twitter datasets, then we ensemble the predictions as a final predictive

<sup>2</sup> <http://scikit-learn.org/>

<sup>3</sup> <http://www.nltk.org/>

model of the cross-genre author profiling task. To investigate the performance of our approach for the task of cross-genre age and gender prediction, we conducted three sets of experiments. We use the blog, social media and Twitter datasets and use the pre-trained models of two sources to test on the remaining source. The results indicate that our approach can be used for the cross-genre author profiling task, where results are better than or equal to the baseline (see Table 3). However, since users’ language in Twitter is different from their language in generating blog posts, in cross-genre author profiling, selecting the training examples from the most similar datasets would be an advantage. However, for PAN2016, since the genre of the test set was hidden, we combine all the available datasets in our submitted software. The results of our submission for PAN2016 on a hidden test data which are evaluated using TIRA [6] are presented in [1].

Table 3: Accuracy of the age and gender prediction using multi-genre ensemble (M-G) approach. Values in bold are higher than the majority baseline (base).

Test (genre)	Train (genres)	English		Spanish					
		Gender	Age	Gender	Age				
blog	Twitter+social media	0.50	<b>0.62</b>	0.37	<b>0.46</b>	0.50	0.50	0.29	<b>0.45</b>
social media	Twitter+blog	0.50	0.50	0.27	<b>0.29</b>	0.50	<b>0.52</b>	0.25	0.25
Twitter	social media+blog	0.50	<b>0.51</b>	0.35	0.35	0.50	<b>0.55</b>	0.32	<b>0.46</b>

### 3 Conclusion

In this paper, we briefly explained our proposed two-level ensemble approach to tackle the cross-genre author profiling task. Our proposed approach is flexible and can incorporate many feature sets and sources of information that are available which makes our approach suitable for the cross-genre author profiling task, where no/little training example is available from the same genre. Experimental results on various datasets and languages indicate the capability of our approach. In our approach, we assigned uniform weights to ensemble the predictive models. However, giving higher weights to the predictive models with better performance may improve the overall performance which is an open path to explore in the future.

### References

1. F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein, “Overview of the 4th Author Profiling Task at PAN 2016: Cross-genre Evaluations,” in *Proc. of the CLEF Evaluation Labs and Workshop*, 2016.
2. F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, “Overview of the 3rd author profiling task at pan 2015,” in *Proc. of the CLEF Evaluation Labs and Workshop*, 2015.
3. G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock, “Computational personality recognition in social media,” *User Modeling and User-Adapted Interaction*, vol. 26, no. 2, pp. 109–142, 2016.
4. J. Villena Román and J.-C. González Cristóbal, “DAEDALUS at PAN 2014: Guessing tweet author’s gender and age,” in *Proc. of the CLEF Evaluation Labs and Workshop*, 2014.
5. S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, “Gender, genre, and writing style in formal written texts,” *TEXT*, vol. 23, no. 3, pp. 321–346, 2003.
6. M. Potthast, T. Gollub, F. Rangel, P. Rosso, E. Stamatatos, and B. Stein, “Improving the Reproducibility of PAN’s Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling,” in *Proc. of the CLEF Evaluation Labs and Workshop*, 2014.