CAPS: A Cross-genre Author Profiling System

Notebook for PAN at CLEF 2016

Ivan Bilan[1], Desislava Zhekova[1]

[1] Center for Information and Language Processing, LMU Munich, Germany

ivan.bilan@gmx.de

zhekova@cis.uni-muenchen.de

Abstract. This paper describes the participation of the Cross-genre Author Profiling System (CAPS) in the PAN16 shared task [15]. The classification system considers parts-of-speech, collocations, connective words and various other stylometric features to differentiate between the writing styles of male and female authors as well as between different age groups. The system achieves the second best score – 74.36% accuracy (with the best performing system (BPS) reaching 75.64%) for gender identification on the official test set (test set 2) for English. Further, for age classification, we report accuracy of 44.87% (BPS: 58.97%). For Spanish, CAPS reaches performance of 62.50% (BPS: 73.21%) for gender and 46.43% (BPS: 51.79) for age, while for Dutch, the accuracy for gender (the task did not target age) is lowest – 55.00% (BPS: 61.80%). For comparison, we also tested CAPS on single genre classification of author gender and age on the PAN14 and PAN15 datasets achieving comparable performance.

Keywords: author profiling, cross-genre, machine learning

1 Introduction

Author profiling is the process of analyzing the sociolect aspect of the writer of an anonymous text sample. The classification tasks in this field range from revealing the author's age and gender to determining the native language or even personality traits of the writer [12]. With the advent of social media websites, the Internet became a treasure trove for classification tasks, similar to author profiling, as the amount of available data increased immensely. The authors often disclose their sociodemographic attributes in their online profiles, although sometimes the information is difficult to trust. For this reason, manually labeling the data is the most appropriate way of collecting a trustworthy dataset. Moreover, since the data is user-generated, it is often freely available in many diverse languages.

Profiling social media authors may be utilized for various purposes. It can, for instance, be used for user-based analysis to find out the demographic characteristics of an average consumer or for targeted advertising to place an ad, tailored to the user's age and gender, next to their written online text. Apart from online texts, profiling the author may also be useful in the field of forensics, for example, to analyze a ransom note.

Previous research into author profiling always concentrated on a single genre. A new view on this problem was proposed by the "Plagiarism Analysis, Authorship

Identification, and Near-Duplicate Detection" (PAN) 2016 [15] (further referred to as PAN16) shared task[1], namely the differentiation among stylistic and structural differences between author classes in one text genre when these differences are observed, analyzed and learned from another text genre.

In this paper, we present the participation of the Cross-genre Author Profiling System (CAPS) in the PAN16 shared task on cross-genre author profiling. Further, in section 2, we provide an overview of relevant work and in section 3 we describe the experimental setup we used. In section 4, we present the results CAPS achieved at the PAN16 task, as well as its evaluation on the PAN14 and PAN15 datasets. Finally, in section 5, we conclude our findings.

2 Related Work

The work by Koppel et al. [8] may be considered the pioneering work into the area of author profiling. With the help of various stylometric features, the authors showed that the gender classification of an author is possible with the aid of machine learning, achieving about 80% accuracy on a small dataset of fiction and non-fiction text samples. Schler et al. [17] explored both gender and age profiling on a dataset of over 70.000 blog posts, subsequently reaching approximately 80% and 75% accuracy for gender and age classification respectively.

Since 2013, the yearly PAN shared task includes author profiling as one of its subtasks. PAN13 [14], conducted in 2013, concentrated on a social media dataset collected from the Netlog[2] website consisting of blogs and including gender and age labels. The task spanned English and Spanish text samples. The best performance for gender profiling reached 59% approximate accuracy for English, whereas the best result for age classification achieved around 66% accuracy.

The following year PAN14 [13] included a wider range of genres, such as tweets collected from Twitter[3], blogs scraped from LinkedIn[4], hotel reviews gathered from the TripAdvisor[5] website and also included a refined version of the PAN13 dataset. In addition to English and Spanish, Dutch and Italian were also targeted by the task. The best result for gender classification in English was achieved on a Twitter dataset with about 76% accuracy. Other genres showed much lower results for gender identification ranging from 53% accuracy for the social media dataset to 73% accuracy for the classification of hotel reviews for English. Age identification proved to be an even more complex task with the best accuracies being between 35% and 50% for various genres in English.

PAN15 [12] concentrated on Twitter samples only and expanded the profiling task to personality traits. The best performance on the PAN15 shared task for both gender and age identification was about 84% and 79% for gender and age classification respectively for the English language. This is significant performance improvement considering the results reached at PAN13.

[1] http://pan.webis.de/clef16/pan16-web/author-profiling.html

[2] http://www.netlog.com

[3] https://twitter.com

[4] https://www.linkedin.com

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

The PAN16 author profiling task concentrates on cross-genre gender and age classification, implying the use of one genre for training and an unseen genre to test the classification model. Such modification of the task can be helpful for underrepresented genres for which no or only small amount of training data is available. However, this modification also increases the complexity of the author profiling task immensely. Not only would a system need to be developed in a general enough manner, but also, to achieve good results, the training set needs to be as similar or close to the test dataset genre as possible.

Our work further demonstrates the discrepancies between the datasets as well as the need to use similar text genres for both training/development and the testing phase. In the following section, we describe our experimental setup and the CAPS system with which we participated in the PAN16 shared task.

3 Experimental Setup

3.1 Workflow Overview

CAPS includes the following main pipeline processes: preprocessing, TF-IDF representation, topic modeling, chi-square term extraction and custom feature extraction. Figure 1 gives a visual representation of the classification pipeline.

3.2 Data Preprocessing and Feature Extraction

Preprocessing: The HTML content and the Bulletin Board Code present in the data are removed as a first preprocessing step. Furthermore, all links are normalized to the special token *[URL]*. Additionally, all user mentions of the form *@username* are translated to *[USER]*. Since the dataset included a considerable number of duplicate text samples, probably due to its automatic collection with the help of web scraping, we excluded all duplicate samples. Table 1 shows the training dataset distribution into each gender and age class as well as a detailed breakdown of the number of authors and the underlying text samples after all duplicate text samples have been discarded. Lemmas and part-of-speech (POS) tags are produced using the TreeTagger [18].

Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF is a procedure widely used in information retrieval and data mining to measure the importance of each word in a corpus which was first formulated in [19]. For the task of cross-genre author profiling a TF-IDF implementation of the scikit-learn machine learning toolkit [11] is used to convert the lemmas and POS-tags, as well as, categorical character n-grams to a matrix of TF-IDF features. The latter is a subdivision of character n-grams into finer classes first introduced by Sapkota et al. [16] and successfully used for the task of author profiling by Maharjan et al. [9].

At the early stage of model training, grid search 5-fold cross-validation was performed for TF-IDF optimization. The TF-IDF vector for the lemma representation shows best results when considering unigrams, bigrams, and trigrams. The part-of-speech TF-IDF vector additionally considers four-gram POS sequences. The categorical character TF-IDF representation considers only trigram characters.

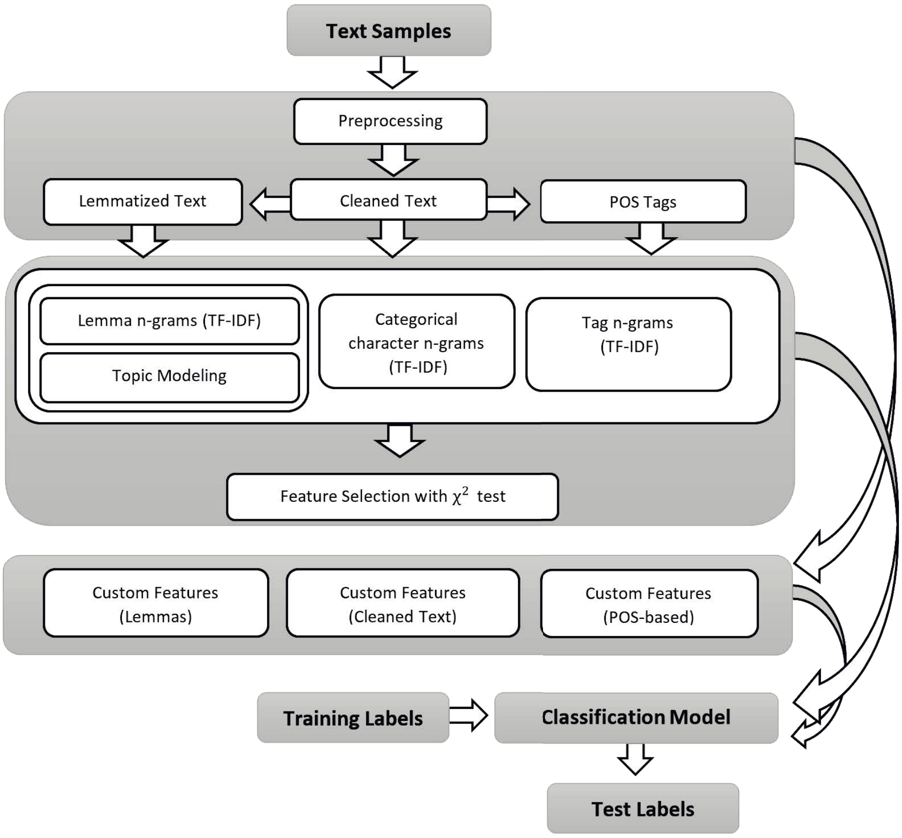

Fig. 1. Classification workflow

Table 1. PAN16 training dataset breakdown (after duplicate removal)

Language		Text Samples					Unique Authors				
English	***Age***	*18-24*	*25-34*	*35-49*	*50-64*	*65-xx*	*18-24*	*25-34*	*35-49*	*50-64*	*65-xx*
	Samples	15725	68936	79338	34668	1435	28	137	181	80	6
	Gender	*Male*			*Female*		*Male*			*Female*	
	Samples	111030			89072		216			216	
	Total	*200102 Text Examples*					*432 Authors*				
Spanish	***Age***	*18-24*	*25-34*	*35-49*	*50-64*	*65-xx*	*18-24*	*25-34*	*35-49*	*50-64*	*65-xx*
	Samples	7146	30730	66287	21449	2869	16	63	38	20	6
	Gender	*Male*			*Female*		*Male*			*Female*	
	Samples	70129			58352		124			125	
	Total	*128481 Text Examples*					*249 Authors*				
Dutch	***Gender***	*Male*			*Female*		*Male*			*Female*	
	Samples	33111			33773		188			191	
	Total	*66884 Text Examples*					*379 Authors*				

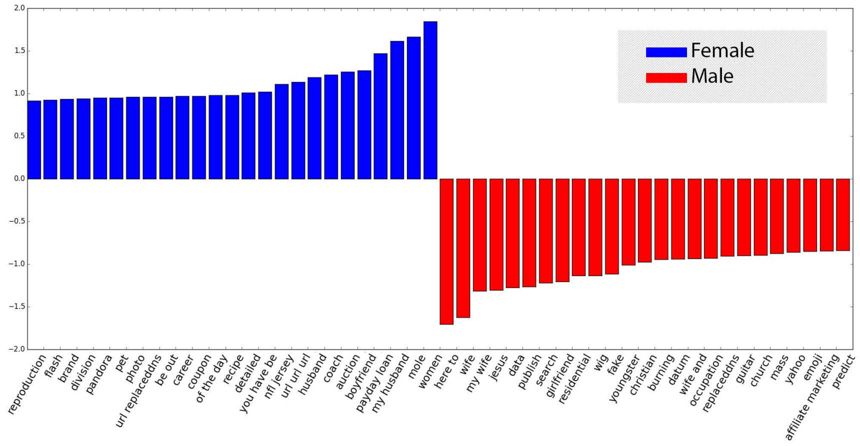

Fig. 2. Correlation coefficient of the 25 most informative lemma n-grams by gender

Fig. 2 shows the most informative lemma, uni-, bi- and trigrams used by male and female authors. The representation is based on the PAN16 Twitter dataset merged with the PAN14 training datasets. Fig. 2 indicates that mentioning the opposite gender provides reliable information about the gender of the author, suggested by the presence of lemmas/n-grams, such as *male*, *my husband*, *boyfriend*, *husband* (used by female authors) and *women*, *wife*, *my wife*, *girlfriend* (used by male authors) amongst the most informative lemmas and bigrams.

Topic Modeling: We apply Latent Dirichlet Allocation (LDA) for topic modeling of the lemmas. LDA is a generative statistical model that assigns probability weights to words and according to those probabilities, assigns the word to a certain automatically generated topic [2].

Chi-Square Term Selection: Since the use of TF-IDF vector representations for the lemma, part-of-speech and character n-grams produces exceedingly high dimensional vectors, for the final classification model only a smaller percentile of the features is selected. For this purpose, the chi-square test is used which tests the independence of term and class occurrences. Subsequently, the terms that are independent of the class are eliminated. This technique ensures the selection of class-dependent features and shrinks the number of the latter, allowing custom features to influence the classification significantly more.

3.3 Custom Features

Apart from TF-IDF vector representation and topic modeling about 40 additional custom features have been developed for age and gender cross-genre profiling. Most of these are stylometric features representing the linguistic style of the author. The features are grouped into four clusters: dictionary-based, POS-based, text structure and stylistic.

Dictionary-Based: The purpose of this set of features is to check whether the raw text representation includes words from a predefined list of tokens. The feature cluster

Table 2. Dictionary-based features

Feature Cluster	Feature Name	Feature Value Examples		
Dictionary-based		**English**	**Spanish**	**Dutch**
	Connective Words	*furthermore, firstly, moreover, hence …*	*pues, como, luego, aunque …*	*zoals, mits, toen, zeker …*
	Emotion Words	*sad, bored, angry, nervous, upset …*	*espanto, carino, calma, peno …*	*boos, moe, zielig, chagrijnig …*
	Contractions	*I'd, let's, I'll, he'd, can't, he'd …*	*al, del, desto, pal', della …*	*m'n, 't, zo'n, a'dam …*
	Familial Words	*wife, husband, gf, bf, mom …*	*esposa, esposo, marido, amiga …*	*vriendin, man, vriend, moeder …*
	Collocations	*dodgy, telly, awesome, freak, troll …*	*no manches, chido, sale …*	*buffelen, geil, dombo, tjo …*
	Abbreviations and Acronyms	*a.m., p.m., Mr., Inc., NASA, asap …*	*art., arch., Avda., Arz., ant. …*	*gesch., geb., nl, notk, mv, vnl …*
	Stop Words	*did, we, ours, you, who, these, because …*	*de, en, que, los, del, donde, como …*	*van, dat, die, was, met, voor …*

consists of dictionaries of connective words, emotion words, contractions, family related words (as proposed by Maharjan et al. [9]), collocations, abbreviations, and acronyms, as well as stop words. All of the dictionaries are adapted for each of the three languages: English, Spanish, and Dutch. A more detailed overview of this cluster with its underlying examples is given in Table 2.

POS-Based: This cluster captures the distribution of the various parts-of-speech. Additionally, this set includes a more complex F-Measure feature, first introduced by Heylighen et al. [7], which indicates how implicit or explicit the text is. The F-Measure is calculated based on the usage of various POS tags in the text.

Text Structure: This feature cluster attempts to analyze the structure of the text and consists of such features as type/token ratio, average word length, and the amount of punctuation signs used in each text sample.

Stylistic: This set of features counts the frequency of use of different adjectival and adverbial suffixes in the text samples. First introduced by Corney et al. [3] for the classification of emails in English, the stylistic feature set is used to represent variation in the writing styles of the authors. For example, men use more emotionally intensive adverbs and adjectives, such as "awful", "dreadfully" or "terribly" [10], which is captured by this feature set.

3.4 Feature Scaling

After all custom feature vectors are extracted, they are scaled. There are several techniques used for feature scaling. Some of the most widely used are normalization and standardization. In the case of cross-genre author profiling, feature values of the training set differ greatly from the feature values of the test set. For example, the lengths of text samples throughout various writing styles range from short, one to two sentence long tweets, to very long samples, as for instance, blog posts. Using normalization or similar form of rescaling would be most suitable to cope with these differences, but the information about outlier values, which can be best captured using standardization, would be lost.

To be able to use standardization for this particular task, a form of feature vector pre-scaling is introduced. Pre-scaling rescales the feature values, which count the number of occurrences of a certain token or a stylistic characteristic, relative to the sample length. A simple solution to this problem could be a division of the feature value by the length of the text in tokens. A more comprehensive approach is presented in Equation 1. It rescales the sample length relative to the lowest mean length of a text sample throughout all possible writing styles that could be represented in both training and test sets and divides the feature value by this rescaled sample length. The rescaled sample length represents the amount of possible smallest sample entities that would fit into the text sample under review. Using this technique, the feature value is always scaled relative to the minimum mean length of all text samples of all represented writing styles. The average length is used instead of minimum or maximum length to better represent the sample length distribution of the writing style that has the shortest text samples on average in the dataset.

$$x^{(i)}_{pre-scaled} = \frac{x^{(i)}}{\left(\frac{len(\varepsilon_i)}{min(\mu y_1 \dots \mu y_n) \mid y_n := len(\varepsilon_{m_1}) \dots len(\varepsilon_{m_n})}\right)} \tag{1}$$

Equation 1 gives a mathematical formulation of the feature pre-scaling approach, where $x^{(i)}$ represents the current feature value, ε_i is the current text sample, μ stands for mathematical mean, y_n represents a genre and ε_{m_n} is the text sample of the genre y_n. *len()* is a function which, given a text sample, returns its length either in tokens or in characters, which makes this interpretation suitable for both types of features that work on the level of tokens and the ones dealing with character representation.

3.5 Classification

There are various approaches to the author profiling classification. The task implies classifying the author of a given text sample, but in many cases, there is a whole set of documents belonging to one author, which raises the question of how to handle the big number of samples per author. It is possible to concatenate all text samples of each author into one uniform sample, as demonstrated by Șulea et al. [20]. Another approach is to build intra-profile relations between text samples and the author profile, as described by Álvarez-Carmona et al. [1], or to classify each text sample separately and then classify the author class based on each text sample belonging to the author. We made use of the latter approach for the development of CAPS.

In fact, some works, including [6], suggest that gender and age classification should be considered as a unified task since these two classes are interrelated. Then, the classification is more accurate when both the gender and age attributes are used simultaneously and not as a separate classification task. Others, like [1], consider it a separate task and build different models for gender and age classification. CAPS also approaches gender and age classification as separate problems. Although it uses the same set of features for both classification models, the classifier used to train the gender profiling model differs from the one used for age identification.

Gender classification is performed using LinearSVC [11], which is based on the LIBLINEAR Classifier [4] and is an implementation of a Support Vector Machine with linear kernel, while age identification makes use of a one-vs.-rest classifier based on Logistic Regression [11]. Various other classifiers have been tested for the task, as for instance, Decision trees classifier or Stochastic Gradient Descent. Our choice of

LinearSVC and Logistic Regression was mainly based on their overall computational efficiency on large datasets.

4 Experimental Results

4.1 PAN16

All of the test datasets used in the PAN shared tasks since 2014 are not openly available and can only be used through the TIRA [5] submission system. TIRA automatically evaluates the classification in terms of accuracy which makes it impossible to review the precision and recall results of the system. For this reason, and since the PAN16 shared task implies testing the dataset on an unseen genre, a smaller subset of the PAN14 hotel reviews and blogs datasets have been used for initial system evaluation in addition to the available PAN16 data.

It is also important to note that the PAN16 task included two different test sets, although no further information is available on them until the final evaluation results are announced. This section reports system results on both these datasets.

Table 3 gives a detailed overview of CAPS' official participation in the PAN16 shared task. CAPS reaches highly competitive performance to the best participating systems in the task, despite the actual complexity of the problem, achieving 74.36% accuracy (on test set 2, which is the official test set used for system ranking) for gender classification on English, closely following the best performing system in this setting that reached 75.64%. However, there is about 20 percent points discrepancy between the system performance on test set 1 (53.74%) and test set 2 (74.36%). This big performance gap is also observed across the performance of the majority of the other participating systems, which shows that the overall performance on this task is highly biased towards the actual test setup. In fact, test set 1 was assembled in a way, such that none of the systems managed to beat the official baseline for this setting – 56.41%. Our assumption was that test set 1 and test set 2 differ based on their genre and that test set 1 is a genre that is more dissimilar to the training set than test set 2. However, this is not the case, since after the evaluation phase it was revealed that both datasets partially overlap which can only be the case if they are from the same genre. Yet, further details on the actual datasets would need to be provided (presumably contained in the task overview paper presented at the PAN16 workshop during the CLEF Conference in September 2016) in order to be able to understand this difference better.

The results for age classification reached only 44.87% for English which reflects the complexity of author profiling when five age groups are targeted. Additionally,

Table 3. Final PAN16 results for CAPS measured in Accuracy

Language (Setting)	Test Set 1	Test Set 2	Average	Baseline
English (Gender)	53.74	74.36	64.05	56.41
English (Age)	29.02	44.87	36.95	19.23
Spanish (Gender)	56.25	62.50	59.38	50.00
Spanish (Age)	23.44	46.43	34.94	17.86
Dutch (Gender)	54.00	55.00	54.50	53.00

the change of genre also poses a difficulty to this task. While the age indicators for one group should stay consistent within the same genre, a different genre might pose a change in the style of writing within the same age group. Specifically, within twitter data, younger age groups tend to make increased use of acronyms, abbreviations, special symbols, etc. which is learned during training. Different genres (expected during testing in this particular task), such as blogs or reviews directly pose a limitation on the use of such an excessive amount of Twitter-specific style as there is no text length limit and in general, all age groups tend to write closer to the standard language. Such a change would be particularly hard to capture within a classification approach. One remedy for this problem could be not to use a single genre for training, but to look into a range of genres that would represent well the different writing styles one particular age group could have.

Apart from the overall lower results reached on the age classification subtask, once again a significant discrepancy between the performance on the two test sets is observed with accuracy scores of 29.02% and 44.87% for each set respectively (an issue co-occurring across all participating systems).

CAPS' results on Spanish and Dutch are much lower, which is easily explained by the fact that the main focus during system development and training lied on English. For Spanish (gender), CAPS achieves 62.50% on test set 2, while for Spanish (age), similar to English, the result is considerably lower – 46.43%. The latter numbers show that Spanish follows the general tendencies observed for English with large discrepancies in performance between test set 1 and 2 and with test set 1 leading to lower performance. For Dutch (gender (age was not targeted during the task)), however, the situation is slightly different – the gap between the performance on both test sets is only 1 percent point, which is considerably lower than the gap observed for English (20.62%) and Spanish (6.25%) for the same setting. The latter is also observed among all other systems. These changes seem to correlate with the size of the datasets available which presumably is also an indicator that once gender and age indicators are well learned for one genre, the change in genre only leads to higher error rates based on the discrepancies of the writing styles between the genders and age groups across the different domains.

4.2 PAN14 and PAN15

For comparison purposes and in addition to the participation in the PAN16 shared task, CAPS was also trained and tested on the datasets of the PAN14 and PAN15 shared tasks through the TIRA evaluation system.

Within the PAN15 setup, CAPS successfully performed on the single genre datasets and produced results also highly competitive to the state-of-the-art systems presented in the task. For instance, CAPS achieved 81.69% for gender profiling on the English dataset, which is only 4.23 percent points lower than the best system presented in that shared task. Table 4 compares the results of CAPS with the best results achieved for gender and age classification in the PAN15 shared task across all three languages (English, Spanish, and Dutch). For CAPS, as for the PAN16 system participation, we provide detailed numbers for both test sets separately, as well as an averaged accuracy score.

Table 4. Results on the PAN15 Datasets measured in Accuracy

Language (Setting)	CAPS			PAN15 Best	Baseline
	Test Set 1	Test Set 2	Average		
English (Gender)	85.71	81.69	83.70	85.92	50.00
English (Age)	73.81	73.24	73.53	83.80	25.00
Spanish (Gender)	93.33	88.64	90.99	96.59	50.00
Spanish (Age)	66.67	67.05	66.86	79.55	25.00
Dutch (Gender)	80.00	78.13	79.07	96.88	50.00

A further comparison between CAPS' performance and the best system results in the PAN14 is given in Table 5. Due to the limited amount of Random Access Memory on the TIRA Virtual System, not all PAN14 datasets could be evaluated in time. Subsequently, due to time constraints no additional resources have been requested from the organizers.

Altogether, Table 5 demonstrates the system's effectiveness on various single genre classifications. The best result of 71.32% in terms of accuracy is achieved on the English hotel reviews dataset for the gender classification, falling about 1 percent point short of the best system's performance (72.59%).

Overall, the results indicate that the current system may also be used in a single genre setting, especially evident by the results on the PAN15 datasets where the model reaches around 70% of average accuracy on both gender and age classification throughout all three languages.

Table 5. Results on the PAN14 Datasets measured in Accuracy

Language (Setting)	Genre	CAPS			PAN14 Best	Baseline
		Test Set 1	Test Set 2	Average		
English (Gender)	Blogs	58.33	66.67	62.50	67.95	57.69
English (Age)	Blogs	25.00	35.90	30.45	46.15	14.10
English (Gender)	Twitter	63.33	60.39	61.86	73.38	59.74
English (Age)	Twitter	56.67	45.45	51.06	50.65	27.92
English (Gender)	Hotel Reviews	73.78	71.32	72.55	72.59	66.26
English (Age)	Hotel Reviews	37.20	34.77	35.99	35.02	27.53
Spanish (Gender)	Blogs	42.86	42.86	42.86	58.93	53.57
Spanish (Age)	Blogs	35.71	44.64	40.18	48.21	16.07
Spanish (Gender)	Twitter	61.54	56.67	59.11	65.56	47.78
Spanish (Age)	Twitter	46.15	48.89	47.52	61.11	46.67

5 Conclusion

The task of cross-genre author profiling is highly complex in comparison to single genre profiling. For the task presented at PAN16, the classification model needs to be very robust and able to adapt the observations learned from short and usually stylistically and grammatically malformed tweet samples to work on any possible form of text as well as samples of any length, for example, hotel reviews, blogs or any other writing style.

CAPS showed promising results in a cross-genre aspect of author profiling regardless of the complexity of the task. The performance of the system on gender identification reached 74.36% for English, 62.50% for Spanish and 55.00% for Dutch displaying great variation across both datasets and across all three languages. With respect to age, CAPS reached 44.87% for English and 46.43% for Spanish (Dutch was not included in the age setting). These low results do outperform the corresponding baseline, but also show that five age groups are probably a too fine-grained distinction to automatically deal with in addition to the change of genre which also increases the complexity of the task. For comparison, we also evaluated CAPS on single genre classification of author gender and age on the PAN14 and PAN15 datasets demonstrating that our system is highly competitive to the state-of-the-art systems applied across all genres. On the PAN15's English dataset CAPS achieved 81.69% for gender classification with the best PAN15 participating system reaching a performance of 85.92%.

CAPS can be further improved in various ways. Firstly, more attention needs to be paid to Spanish and Dutch since the included custom features are only tailored for English and simply adjusted to function on other languages represented in the shared task. Second, the current model considers each text sample as a separate entity that does not correlate with the other text samples belonging to the author. Some form of text sample-author profile interrelation could improve the model performance.

References

1. Álvarez-Carmona, M., López-Monroy, P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Escalante, H.: INAOE's participation at PAN'15: Author Profiling task - Notebook for PAN at CLEF 2015. In: CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers, Toulouse, France (2015)
2. Coelho, L. P., Richert, W.: Building Machine Learning Systems with Python, Second Edition. Packt Publishing Ltd. (2015)
3. Corney, M., De Vel, O., Anderson, A., Mohay, G.: Gender-preferential text mining of e-mail discourse. In: Proceedings - Annual Computer Security Applications Conference, ACSAC, vol. 2002-January, pp.282–289 (2002)
4. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A Library for Large Linear Classification. The Journal of Machine Learning 9, 1871–1874 (2008)
5. Gollub, T., Stein, B., Burrows, S., Hoppe, D.: TIRA: Configuring, executing, and disseminating information retrieval experiments. Proceedings - International Workshop on Database and Expert Systems Applications, DEXA, 151–155 (2012)
6. González-Gallardo, C., Montes, A., Sierra, J., Núñez-Juárez, J. A., Salinas-López, A. J., Ek, J.: Tweets Classification Using Corpus Dependent Tags, Character and POS N-grams - Notebook for PAN at CLEF 2015. CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers (2015)
7. Heylighen, F., Dewaele, J.: Variation in the Contextuality of Language: An Empirical Measure. Foundations of Science 7(3), 293–340 (2002)
8. Koppel, M., Argamon, S., Shimoni, A.: Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing 17(4), 401–412 (2002)

9. Maharjan, S., Solorio, T.: Using Wide Range of Features for Author Profiling - Notebook for PAN at CLEF 2015. In: CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, Toulouse, France (2015)
10. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. In: Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp.207–217 (2010)
11. 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. Journal of Machine Learning Research 12, 2825–2830 (2011)
12. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, Toulouse, France (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: CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, Sheffield, UK (2014)
14. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers, Valencia, Spain (2013)
15. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Evaluations Concerning Cross-genre Author Profiling. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2016)
16. Sapkota, U., Bethard, S., y Gómez, M. M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), pp.93–102 (2015)
17. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp.199–205 (2006)
18. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, pp.44–49 (1994)
19. Spärck Jones, K.: A Statistical Interpretation of Term Specificity and its Retrieval. Journal of Documentation 28(1), 11–21 (1972)
20. Şulea, O.-M., Dichiu, D.: Automatic Profiling of Twitter Users Based on Their Tweets - Notebook for PAN at CLEF 2015. CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers (2015)